Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design

Aikaterini Alexaki[1,†], Jacob Kames[1,†], David D. Holcomb[1,†], John Athey[1], Luis V. Santana-Quintero[2], Phuc Vihn Nguyen Lam[2], Nobuko Hamasaki-Katagiri[1], Ekaterina Osipova[2], Vahan Simonyan[2], Haim Bar[3], Anton A. Komar[4] and Chava Kimchi-Sarfaty[1,]

*1 - **Division of Plasma Protein Therapeutics,** Office of Tissue and Advanced Therapies, Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993, USA*

*2 - **High Performance Integrated Environment,** Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993, USA*

*3 - **Department of Statistics,** University of Connecticut, Storrs, CT 06268, USA*

*4 - **Center for Gene Regulation in Health and Disease,** Cleveland State University, Cleveland, OH 44115, USA*

***Correspondence to Chava Kimchi-Sarfaty and Anton A. Komar:** Division of Plasma Protein Therapeutics, Office of Tissue and Advanced Therapies, Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993, USA. Center for Gene Regulation in Health and Disease, Cleveland State University, Cleveland, OH 44115, USA Chava.kimchi-sarfaty@fda.hhs.gov, a.komar@csuohio.edu.*

https://doi.org/10.1016/j.jmb.2019.04.021 Division of Plasma Protein Therapeutics, Office of Tissue and Advanced Therapies, Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993, USA.
Edited by Michael Sternberg

Abstract

Usage of sequential codon-pairs is non-random and unique to each species. Codon-pair bias is related to but clearly distinct from individual codon usage bias. Codon-pair bias is thought to affect translational fidelity and efficiency and is presumed to be under the selective pressure. It was suggested that changes in codon-pair utilization may affect human disease more significantly than changes in single codons. Although recombinant gene technologies often take codon-pair usage bias into account, codon-pair usage data/tables are not readily available, thus potentially impeding research efforts. The present computational resource (https://hive.biochemistry.gwu.edu/review/codon2) systematically addresses this issue. Building on our recent HIVE-Codon Usage Tables, we constructed a new database to include genomic codon-pair and dinucleotide statistics of all organisms with sequenced genome, available in the GenBank. We believe that the growing understanding of the importance of codon-pair usage will make this resource an invaluable tool to many researchers in academia and pharmaceutical industry.

Published by Elsevier Ltd.

Introduction

The phenomenon of codon usage bias is well recognized and studied across species [1,2]. Similarly, there is a substantial bias in codon-pair utilization, referred to as bicodon or dicodon bias, or sometimes (more broadly) as codon context [3,4]. Codon-pairs are encountered, in any given genome, at different frequencies than would be expected based on the individual codon usage bias of that genome [5–7]. First described in *Escherichia coli* in 1985 [8], codon-pair usage bias has, since then, been described in all three domains of life: eukarya, archaea, and bacteria [9]. Recently, the importance of alternative use of synonymous codon-pairs was also highlighted in relation to human diseases [10].

Although codon-pair usage bias is a well-recognized phenomenon, it is still unclear to what extent it is shaped by selection or by mutational drift. Several lines of evidence point to structural constraints exerted at the ribosome decoding center that shape the observed codon pair bias. For example, studies in *E. coli* have suggested that the size of the transfer ribonucleic acid (tRNA) variable loop

determines whether the 3′ nucleotide adjacent to the codon has a context effect [11], while others have argued that the P-site wobble position, within the codon–anticodon interaction, and the A-site anticodon loop and acceptor stem have a decisive effect on the observed genomic codon-pair patterns [7]. Recent studies in yeast, which showed that the wobble base pairing has a critical role in whether a codon-pair would have an inhibitory effect on translation [12], are in agreement with these earlier observations. Further supporting the role of codon-pair bias, numerous studies have provided evidence on the effect of codon usage bias on translational fidelity and rate [6,13,14]. On the other hand, there is a substantial body of evidence suggesting that biases in codon and codon-pair usage arise from GC-biased gene conversion [15,16] and that codon-pair bias is directly related to dinucleotide bias [17], highlighting the role of mutational processes. Since codon bias, codon-pair bias, and dinucleotide frequency are inevitably connected, it is often debated whether one is causing the other and what the underlying mechanism for the observed bias is.

Despite an incomplete understanding of these phenomena, tools that have incorporated codon-pair bias in their algorithms for gene optimization or deoptimization have been proven successful. In 2007, Translation Engineering [18] was patented, a methodology aiming to fine-tune translation by accounting for codon-pair usage, translational pausing signals, and RNA secondary structure, in addition to codon usage. Soon after, a simpler method was patented [19], which considered only codon usage and codon-pair usage. Since then, freely available optimization tools have been published, such as EuGene, which also accounts for GC content in addition to codon and codon-pair usage, as well as Codon Optimization OnLine (COOL), a web-based tool [20], and methodologies that also consider host expression profiles [21]. These optimization approaches have been used extensively, exhibiting significant improvements over methods that do not consider codon-pair usage; for example, optimization of interferon gamma in Chinese hamster ovary (CHO) cells displayed on average a 14-fold increase when codon-pair usage was incorporated into the optimization algorithm compared to a 9-fold increase when it was not [22]. Recently, a web-based tool (CCtool) was published that computes the codon-pair bias of a gene in an effort to evaluate the contribution of codon-pair optimization relative to other features such as codon bias and mRNA secondary structure [5,23]. Still, access to codon-pair usage statistics is not readily available, leading to significant work burden every time a new genome is analyzed.

At the other end of the spectrum, codon-pair deoptimization of virus genes has been used widely in vaccine development. Synthetic attenuated virus engineering (SAVE) [24], a strategy for codon-pair deoptimization, recodes a given amino acid sequence utilizing rare bicodons while controlling codon bias and RNA free folding energy. The major advantages of codon-pair based deoptimization strategies over conventional viral attenuation approaches are that (i) it is a systematic method applicable, in theory, to any virus, and (ii) the attenuation is not subject to reversion, simply because of the sheer number of mutations. Genes of the porcine reproductive and respiratory syndrome virus [25–27], influenza virus [28,29], Marek's disease virus [30–32], and Zika virus [33] are among those that have been successfully codon-pair deoptimized and are currently at various stages of vaccine development. Some bacterial genomes have also been partially reengineered through codon-pair deoptimization [34,35], resulting in attenuation of their pathogenicity and thus broadening the applicability of this method. Undoubtedly, having easy access to codon-pair bias statistics would simplify the deoptimization process for vaccine development and other genetic engineering tasks.

In addition to gene design applications, having access to accurate and comprehensive codon-pair usage statistics can facilitate disease prediction caused by synonymous mutations. An increasing number of synonymous mutations have been demonstrated to cause disease and, in many cases, there is no clear mechanism explaining this association. Interruption of splicing signals of precursor mRNAs, disruption of regulatory binding-sites of transcription factors and miRNAs, and modification of the secondary structure of mRNAs are all possible mechanisms. An additional mechanism that has been difficult to study is the potential of a synonymous mutation to cause perturbations in the translational kinetics of the protein leading to altered conformation and function. It has been shown in some systems that rare codons are translated slower than common codons [36,37]; however, ribosome profiling in higher eukaryotes has yielded controversial results [38], generally not supporting the direct relationship between differences in elongation rates and codon usage. Recently, McCarthy *et al*. [39] examined 35 synonymous single nucleotide polymorphisms linked to disease and proposed that codon-pair usage, instead of codon usage, could be responsible for altered translational kinetics. However, since then, more reports have implicated codon usage bias in various diseases including Alzheimer's [40] and autism [41], while other factors such as dinucleotide usage and the strength of codon anticodon interaction could also be involved [42]. Taken together, these somewhat controversial data may suggest that a complex interaction between several factors (codon usage, mRNA secondary structure, the nature of the nascent chain itself, etc.) influence translation speed and, as a result, protein conformation. Although a codon usage database has long been available to facilitate research on the contribution of codon usage on translational kinetics and disease, a codon-pair usage database has been lacking. Presenting codon usage and codon-pair usage statistics within the same

computational resource could thus greatly facilitate predicting the propensity of synonymous mutations to cause disease.

The genetic code is inherently complex with several levels of organization built into it. Chromatin structure, transcriptional regulation, splicing motifs, microRNA binding, mRNA stability, translational efficiency, and cotranslational folding may all have a role in determining the nucleotide sequence by applying selective pressure [43]. At the same time, mutational drift may counteract the effect of selection processes. Having accurate GC content, dinucleotide, codon, and codon-pair statistics for essentially all species with available sequence data can be instrumental in numerous areas of research. Data mining from this single comprehensive, up-to-date computational resource can (i) shed light on the complex relationship between selective pressure and mutational processes that is shaping any genome, (ii) enable the generation of better algorithms for gene optimization and deoptimization, and (iii) facilitate disease prediction for both synonymous and non-synonymous mutations.

Results and Discussion

We have created codon and codon-pair usage tables (CoCoPUTs), a new, regularly updated website that encompasses the previously generated High-Performance Integrated Virtual Environment codon usage tables (HIVE-CUTs) [1] but has also been expanded to include codon-pair usage and dinucleotide statistics. On the CoCoPUTs homepage, there is a *service help* tab, containing information about each of the functions and outputs of this resource, and a user-friendly search engine where the user can look for species or a taxonomic rank of interest. Once a search is submitted, the results appear in several tabs. The *codon-pair heatmap*, *codon-pair data*, and *dinucleotide frequency* are newly available tabs. The *dinucleotide frequency* tab includes both total dinucleotides and junction dinucleotides, which refer to the dinucleotides bridging two codons, as well as comparison between these values. To our knowledge, there is no other computational resource that provides comprehensive codon-pair and dinucleotide statistics. *Codon bias* and *files to download* are tabs that existed in the HIVE-CUTs website and have now been expanded to include codon-pair usage bias metrics and codon-pair usage files, respectively. *Taxonomy*, *codon usage table*, *GC%* (GC content expressed as a percentage of total), and *codon frequencies* tabs have been maintained from the HIVE-CUTs website. All data have been recalculated with the most recent available versions of GenBank and RefSeq.

There are 4096 possible codon-pairs, 192 of which are stop codon-pairs (having a stop as their second codon) and another 192 are non-sense codon-pairs [having a stop as their first codon and should not be found within the coding sequences (CDSs)]. To enable simultaneous visualization of all codon-pair relative usage, a heatmap representation was chosen. The heatmaps display codon-pair data by percentile rank. This allows users to compare codon-pairs more easily than codon-pair frequencies, wherein many codon-pairs would be similarly colored, except for a few outliers. Selecting any codon-pair on the heatmap opens a window with the percentile rank of that codon-pair. As an example, Fig. 1 demonstrates the *Homo sapiens* codon-pair percentile heatmap and the percentile rank of CTGGAG. At the heatmap presentation, abundant codon-pairs are seen in red, while rare codon-pairs are visible in blue. Unsurprisingly, the most common codon GAG is part of the most common codon-pair GAGGAG; however, GAA, also a very common codon (that just as GAG is translated to glutamate) is not encountered frequently in tandem with GAG but is very often found in tandem by itself. Conversely (when stop codons and codon-pairs are taken out of the percentile ranking), the rarest codon-pair, CGTACG, does not contain the rarest codon, TCG; furthermore, TCGTCG is not one of the rarest codons (91st rarest), illustrating that codon bias and codon-pair bias often differ. The *codon-pair data* are also available in a separate tab in plain text format with the raw total counts for each codon-pair in the genome. When a search for multiple species is conducted, a heatmap is generated for each species, while data from all species are combined in the plain text *codon-pair data* tab. Stop and non-sense codon-pairs are included in the plain text format but not in the heatmaps. Searches for taxonomical ranks of any level are also possible and generate the same type of output. The complete data set can be downloaded in delimited text format (from the *files to download* tab) for further analysis. The files contain codon, codon pair, dinucleotide, and junction dinucleotide counts for all available species, as well as codon counts from all CDSs. Data derived from RefSeq and GenBank are in separate files.

To compare the extent of codon and codon-pair usage bias in a given species or taxonomic rank, we have calculated the effective number of codon-pairs (ENcp) as a metric of codon-pair usage bias. Similarly to the effective number of codons (ENc), a metric to measure codon usage bias in a genome based on deviation from an expected equal usage of synonymous codons [44], ENcp can range from 20 (very biased) to 61 (not biased at all). Both metrics can be found in the *effective number* tab. More information on this feature can be found on the service help of the website.

Since it remains an open question whether codon-pair bias is a result of dinucleotide bias [15,45], dinucleotide frequencies were also calculated and presented (Fig. 2). Codon-pair bias is mostly associated with dinucleotides at codon junctions,

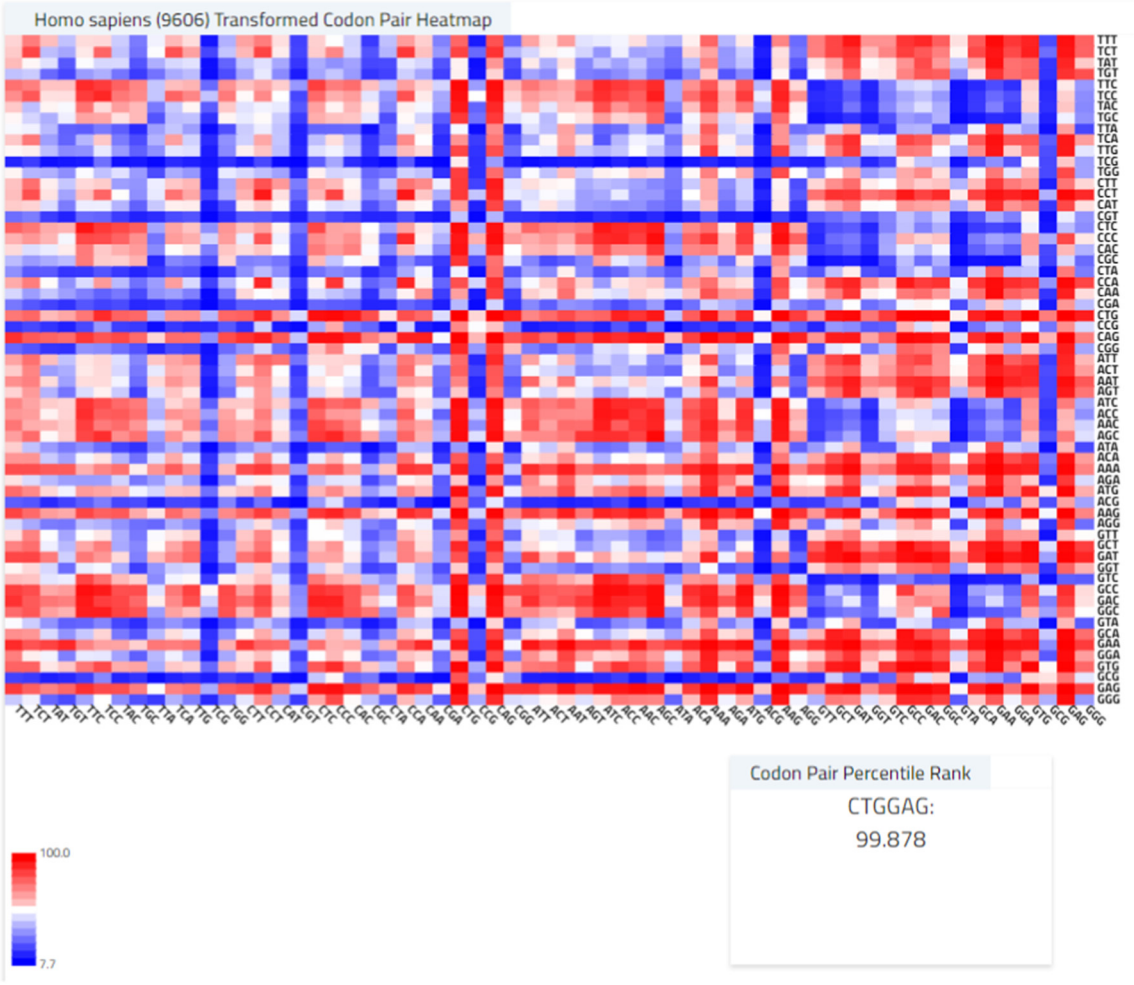

Fig. 1. Screenshot of the *H. sapiens* codon-pair usage heatmap. The first codon of each pair is given on the *x*-axis, and the second codon is given on the *y*-axis. Codon pair frequencies are transformed into percentile ranks. The percentile rank for CTGGAG is shown.

and since the count of these dinucleotides cannot be derived from codon usage statistics, separate counts were generated for total dinucleotides and junction dinucleotides. To enable comparisons between different organisms and clades, multiple queries can be submitted simultaneously, and their results are plotted in combination. For each species, the total dinucleotides and junction dinucleotides are plotted in the same graph for comparison.

Figure 2A shows the total dinucleotide frequency for *H. sapiens*, Zika virus (*Flavivirus flaviviridae*), and its arthropod vector, *Aedes aegypti.* Fig. 2B presents only the junction dinucleotide frequency of these species. Strikingly, the frequency of the CG dinucleotide in Zika virus is much more similar to *H. sapiens* than to *A. aegypti*. Figure 2C shows the dinucleotide frequency and junction dinucleotide frequency in *A. aegypti*. Clearly, some dinucleotides such as CA, CG, AT, AC, AA, and GG occur at much different frequencies, dependent on their position. These three graphs would be automatically generated and presented in the *dinucleotide frequency* tabs following a search for *H. sapiens*, Zika virus, and *A. aegypti.*

Combining numerous statistics and information of each genome in one online computational tool facilitates novel inquiries and data mining. For example, the similarity of the Zika virus genome to its hosts' genomes, *H. sapiens* and *A. aegypti,* can be interrogated through its codon usage, codon-pair usage, and dinucleotide frequency (Table 1). When considering relative synonymous codon usage (RSCU) of zika virus, *H. sapiens*, and *A. aegypti*, it appears that the Zika virus genome is more similar to the *H. sapiens* genome and one may assume that the virus has adapted for optimal replication in human cells. However, the observed/expected codon-pair frequency ratios suggest that the *H. sapiens* and the *A. aegypti* genomes are relatively similar to each other but very different from the Zika virus genome. In contrast, when considering the dinucleotide and

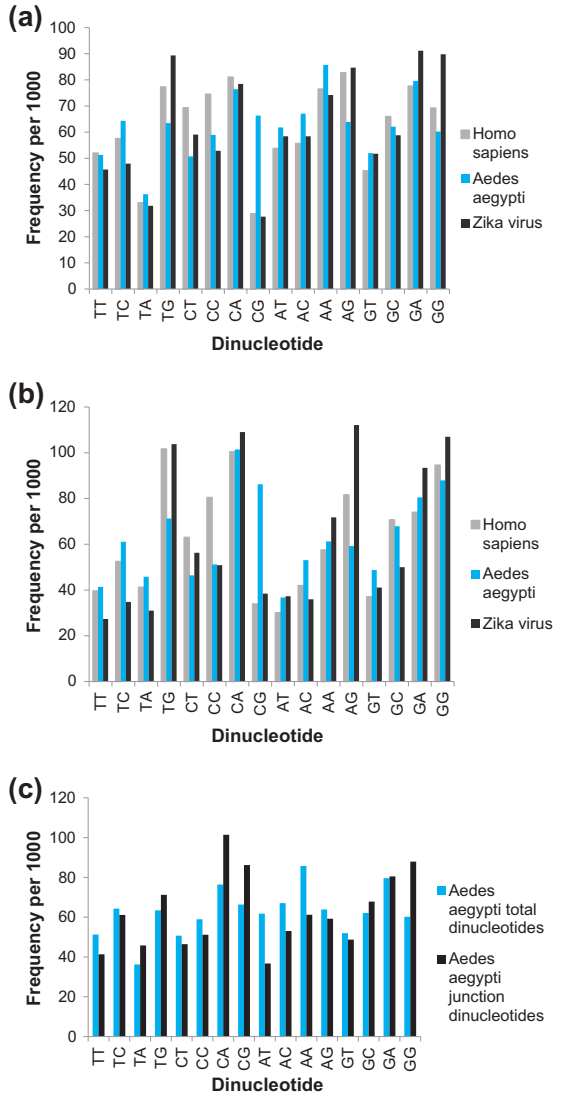

Fig. 2. Total dinucleotide frequencies of *H. sapiens*, *A. aegypti*, and Zika virus CDS (A). Junction dinucleotide frequencies for *H. sapiens, A. aegypti*, and Zika virus CDS (B). Total and juction nucleotide frequencies of *A. aegypti* (C).

junction dinucleotide frequencies, *H. sapiens* and *A. aegypti* appear quite distant. Clearly, to gain a holistic understanding of the similarities of these or any other genomes, each parameter must be weighed in. Recently, a machine learning-based method was developed to predict animal reservoirs and arthropod vectors directly from viral genome sequences by examining codon-pair, dinucleotide, codon, and amino acid biases, recognizing that accounting for all these biases that are present in the genome adds power to their method [46]. Efforts such as this would greatly benefit from the availability of the CoCoPUTs computational resource. A related open question [17,45,47,48] is whether codon-pair bias or dinucleotide bias is responsible for the attenuation effect that is being observed when codon-pair bias deoptimization is applied in virus genomes. Clearly, the availability of codon-pair, dinucleotide, and junction dinucleotide statistics from a wide range of viruses and their hosts will enable the indisputable resolution of this debate.

In conclusion, the presented computational resource will facilitate research in the areas of (i) gene optimization and viral genome deoptimization for vaccine development, (ii) analysis of translational kinetics and of host-pathogen co-evolutionary relationships, and (iii) predicting the impact of single nucleotide polymorphism in disease. The growing understanding of the importance of codon-pair usage will potentially lead to new areas where this resource may be useful.

Methods

Input data

Codon usage, codon-pair usage, and dinucleotide usage were computed for all available species in GenBank (Release 230, February 15, 2019) databases and for all RefSeq assemblies available as of March 13, 2019. Data were acquired as described by Athey *et al*. [1]. RefSeq data account for 155,710 assemblies and GenBank data represent 1,275,531 individual species. CoCoPUTs will be updated regularly whenever a new GenBank release is available, and each version of the database will remain available to provide a stable reference.

Data processing

Input data were processed using C++98 with compiler G++4.8.5 and custom libraries to parse annotations from input files. All protein coding

Table 1. Pearson correlations of codon bias, codon-pair bias, and dinucleotide frequency between *H. sapiens, A. aegypti*, and Zika virus

	RSCU	Observed/expected codon-pair frequency	Dinucleotide and junction dinucleotide frequencies
A. aegypti–H. sapiens	0.81746	0.81054	0.59227
A. aegypti–Zika virus	0.77633	0.4073	0.61658
H. sapiens–Zika virus	0.91298	0.51561	0.86557

RSCU and observed/expected codon-pair frequency are used as measures of codon and codon-pair bias, respectively.

sequences with "CDS" tags were parsed for codon, codon-pair, and dinucleotide usage. All CDSs that were tagged as "low-quality" protein or "pseudogene" were excluded from the computation. Furthermore, any CDSs with improper annotations were excluded. TaxID numbers and scientific names of organisms were parsed as previously described [1]. The overall number of excluded sequences is very low and has a negligible impact on overall codon, codon-pair, and dinucleotide usage calculations. Individual codons, codon-pairs, and dinucleotides containing ambiguous nucleotides were excluded from the calculation without affecting the inclusion of the remaining CDSs. Execution of the entire pipeline (data acquisition, processing, and output) was performed using FDA's HIVE [49].

Output and interface

Codon, codon-pair, and dinucleotide usage data are organized by assembly accession number for RefSeq entries or by species name for GenBank entries, and the user may retrieve data according to organelle as described by Athey *et al*. [1]. Data derived from RefSeq assemblies are preferentially displayed in searches such that data derived from GenBank entries will only be displayed when no RefSeq assembly was available for that species.

The infrastructure of the website interface has been previously described [1]. In addition to graphical display of codon usage frequencies, total dinucleotide frequencies, junction dinucleotide frequencies, and the percentile heatmap visualization of codon-pair usage are accessible through separate tabs on the interface. Percentile rank of individual codon pairs may be visualized by selecting the codon pair of interest on the interactive heatmap, while the raw data appear in the *codon pair data* tab. The ENc and ENcp are accessible together at the *Effective Number* tab. The ENc is a metric to measure codon usage bias based on deviation from an expected equal usage of synonymous codons [44]. The ENcp is a metric to measure codon-pair usage bias, defined analogously to ENcs, with the addition of a square root. $\widehat{N_{\mathrm{cp}}} = \sqrt{\sum_m \left({}^{k_m}/_{\widehat{F_m}} \right)}$, where $\widehat{F_m}$ is the average degeneracy of all amino acid pairs with m synonymous codon pairs, and k_m is the number of amino acid pairs with m synonymous representations. For standard genetic code, $\widehat{N_{\mathrm{cp}}} = \sqrt{4 + \frac{36}{\widehat{F_2}} + \frac{4}{\widehat{F_3}} + \frac{101}{\widehat{F_4}} + \frac{30}{\widehat{F_6}} + \frac{90}{\widehat{F_8}} + \frac{1}{\widehat{F_9}} + \frac{64}{\widehat{F_{12}}} + \frac{25}{\widehat{F_{16}}} + \frac{6}{\widehat{F_{18}}} + \frac{30}{\widehat{F_{24}}} + \frac{9}{\widehat{F_{36}}}}$. The calculation is performed for many different genetic codes, and users should identify the genetic code of the organism they are interested in.

All output data files can be downloaded by the user in a delimited text format.

RSCU (Table 1) is calculated as in Sharp *et al*. [50].

For codon pairs, we calculate the observed/expected frequency ratio (Table 1) by dividing the codon pair frequency by the frequencies of the constituent codons. For a codon pair ABCDEF, that is $\frac{F(\mathrm{ABCDEF})}{F(\mathrm{ABC})\,F(\mathrm{DEF})}$.

Acknowledgments

This work was supported by funds from the Hemostasis Branch/Division of Plasma Protein Therapeutics/Office of Tissues and Advanced Therapies/Center for Biologics Evaluation and Research of the U.S. Food and Drug Administration and in part by an appointment to the Research Participation Program at the Center for Biologics Evaluation and Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration (C.K.-S.).

Received 1 February 2019;
Received in revised form 10 April 2019;
Accepted 15 April 2019
Available online 26 April 2019

Keywords:
codon-pair bias;
codon context;
dinucleotide frequency;
gene optimization/deoptimization;
GenBank

†A.A., J.K., and D.D.H. contributed equally to this work.

Abbreviations used:
CoCoPUTs, Codon and codon-pair usage tables; HIVE, High-performance Integrated Virtual Environment; HIVE-CUTs, HIVE codon usage tables; CDS, Coding sequence; RSCU, relative synonymous codon usage; ENc, effective number of codons; ENcp, effective number of codon-pairs.

References

[1] J. Athey, A. Alexaki, E. Osipova, A. Rostovtsev, L.V. Santana-Quintero, U. Katneni, V. Simonyan, C. Kimchi-Sarfaty, A new and updated resource for codon usage tables, BMC Bioinform 18 (2017) 391.

[2] A.A. Komar, The Yin and Yang of codon usage, Hum. Mol. Genet. 25 (2016) R77–R85.
[3] R.H. Buckingham, Codon context, Experientia 46 (1990) 1126–1133.
[4] R.H. Buckingham, Codon context and protein synthesis: enhancements of the genetic code, Biochimie 76 (1994) 351–354.
[5] D. Papamichail, H.M. Liu, V. Machado, N. Gould, J.R. Coleman, G. Papamichail, Codon context optimization in synthetic gene design, IEEE-ACM Trans. Comput. Biol. Bioinform. 15 (2018) 452–459.
[6] B. Irwin, J.D. Heck, G.W. Hatfield, Codon pair utilization biases influence translational elongation step times, J. Biol. Chem. 270 (1995) 22801–22806.
[7] J.R. Buchan, L.S. Aucott, I. Stansfield, tRNA properties help shape codon pair preferences in open reading frames, Nucleic Acids Res. 34 (2006) 1015–1027.
[8] M. Yarus, L.S. Folley, Sense codons are found in specific contexts, J. Mol. Biol. 182 (1985) 529–540.
[9] A. Tats, T. Tenson, M. Remm, Preferred and avoided codon pairs in three domains of life, BMC Genomics 9 (2008) 463.
[10] L.A. Diambra, Differential bicodon usage in lowly and highly abundant proteins, PeerJ 5 (2017), e3081.
[11] J.F. Curran, E.S. Poole, W.P. Tate, B.L. Gross, Selection of aminoacyl-tRNAs at sense codons: the size of the tRNA variable loop determines whether the immediate 3′ nucleotide to the codon has a context effect, Nucleic Acids Res. 23 (1995) 4104–4108.
[12] C.E. Gamble, C.E. Brule, K.M. Dean, S. Fields, E.J. Grayhack, Adjacent codons act in concert to modulate translation efficiency in yeast, Cell 166 (2016) 679–690.
[13] F.F.V. Chevance, S. Le Guyon, K.T. Hughes, The effects of codon context on in vivo translation speed, PLoS Genet. 10 (2014), e100439.
[14] G.R. Moura, M. Pinheiro, A. Freitas, J.L. Oliveira, J.C. Frommlet, L. Carreto, A.R. Soares, A.R. Bezerra, M.A.S. Santos, Species-specific codon context rules unveil non-neutrality effects of synonymous mutations, PLoS One 6 (2011), e26817.
[15] F. Pouyet, D. Mouchiroud, L. Duret, M. Semon, Recombination, meiotic expression and human codon usage, Elife 6 (2017), e27344.
[16] P. Mazumdar, R.B. Othman, K. Mebus, N. Ramakrishnan, J.A. Harikrishna, Codon usage and codon pair patterns in non-grass monocot genomes, Ann. Bot. 120 (2017) 893–909.
[17] D. Kunec, N. Osterrieder, Codon pair bias is a direct consequence of dinucleotide bias, Cell Rep. 14 (2016) 55–67.
[18] G.W. Hatfield, D.A. Roth, Optimizing scaleup yield for protein production: computationally optimized DNA assembly (CODA) and translation engineering((TM)), Biotechnol. Annu. Rev. 13 (13) (2007) 27–42.
[19] J.A. Roubos, N.N.M.E. Van Peij, Method for Achieving Improved Polypeptide Expression, DSM IP Assets BV, USA, 2008.
[20] J.X. Chin, B.K.S. Chung, D.Y. Lee, Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design, Bioinformatics 30 (2014) 2210–2212.
[21] A.M. Lanza, K.A. Curran, L.G. Rey, H.S. Alper, A condition-specific codon optimization approach for improved heterologous gene expression in Saccharomyces cerevisiae, BMC Syst. Biol. 8 (2014) 33.
[22] B.K.S. Chung, F.N.K. Yusufi, Mariati, Y.S. Yang, D.Y. Lee, Enhanced expression of codon optimized interferon gamma in CHO cells, J. Biotechnol. 167 (2013) 326–333.
[23] G. Moura, M. Pinheiro, J. Arrais, A.C. Gomes, L. Carreto, A. Freitas, J.L. Oliveira, M.A.S. Santos, Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure, PLoS One 2 (2007).
[24] J.R. Coleman, D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, S. Mueller, Virus attenuation by genome-scale changes in codon pair bias, Science 320 (2008) 1784–1787.
[25] Y.Y. Ni, Z. Zhao, T. Opriessnig, S. Subramaniam, L. Zhou, D.J. Cao, Q. Cao, H.C. Yang, X.J. Meng, Computer-aided codon-pairs deoptimization of the major envelope GP5 gene attenuates porcine reproductive and respiratory syndrome virus, Virology 450 (2014) 132–139.
[26] D. Evenson, P.F. Gerber, C.T. Xiao, P.G. Halbur, C. Wang, D. Tian, Y.Y. Ni, X.J. Meng, T. Opriessnig, A porcine reproductive and respiratory syndrome virus candidate vaccine based on the synthetic attenuated virus engineering approach is attenuated and effective in protecting against homologous virus challenge, Vaccine 34 (2016) 5546–5553.
[27] L. Gao, L.H. Wang, C. Huang, L.L. Yang, X.K. Guo, Z.B. Yu, Y.H. Liu, P. Yang, W.H. Feng, HP-PRRSV is attenuated by de-optimization of codon pair bias in its RNA-dependent RNA polymerase nsp9 gene, Virology 485 (2015) 135–144.
[28] S. Mueller, J.R. Coleman, D. Papamichail, C.B. Ward, A. Nimnual, B. Futcher, S. Skiena, E. Wimmer, Live attenuated influenza virus vaccines by computer-aided rational design, Nat. Biotechnol. 28 (2010) 723–726.
[29] B.S. Kaplan, C.K. Souza, P.C. Gauger, C.B. Stauft, J.R. Coleman, S. Mueller, A.L. Vincent, Vaccination of pigs with a codon-pair bias de-optimized live attenuated influenza vaccine protects from homologous challenge, Vaccine 36 (2018) 1101–1107.
[30] S.J. Conrad, R.F. Silva, C.J. Hearn, M. Climans, J.R. Dunn, Attenuation of Marek's disease virus by codon pair deoptimization of a core gene, Virology 516 (2018) 219–226.
[31] K. Eschke, J. Trimpert, N. Osterrieder, D. Kunec, Attenuation of a very virulent Marek's disease herpesvirus (MDV) by codon pair bias deoptimization, PLoS Pathog. 14 (2018).
[32] P.H. Khedkar, N. Osterrieder, D. Kunec, Codon pair bias deoptimization of the major oncogene meq of a very virulent Marek's disease virus, J. Gen. Virol. 99 (2018) 1705–1716.
[33] P.H. Li, X.L. Ke, T. Wang, Z.Y. Tan, D. Luo, Y.J. Miao, J.H. Sun, Y. Zhang, Y. Liu, Q.X. Hu, F.Q. Xu, H.Z. Wang, Z.H. Zheng, Zika virus attenuation by codon pair deoptimization induces sterilizing immunity in mouse models, J. Virol. 92 (2018).
[34] J.R. Coleman, D. Papamichail, M. Yano, M.D. Garcia-Suarez, L.A. Pirofski, Designed reduction of *Streptococcus pneumoniae* pathogenicity via synthetic changes in virulence factor codon-pair bias, J. Infect. Dis. 203 (2011) 1264–1273.
[35] L.M. Runco, C.B. Stauft, J.R. Coleman, Tailoring the immune response via customization of pathogen gene expression, J. Pathog. 2014 (2014) 651568.
[36] A.A. Komar, T. Lesnik, C. Reiss, Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation, FEBS Lett. 462 (1999) 387–391.
[37] F. Buhr, S. Jha, M. Thommen, J. Mittelstaet, F. Kutz, H. Schwalbe, M.V. Rodnina, A.A. Komar, Synonymous codons direct cotranslational folding toward different protein conformations, Mol. Cell 61 (2016) 341–351.
[38] N.T. Ingolia, Ribosome footprint profiling of translation throughout the genome, Cell 165 (2016) 22–33.
[39] C. McCarthy, A. Carrea, L. Diambra, Bicodon bias can determine the role of synonymous SNPs in human diseases, BMC Genomics 18 (2017).

[40] J.E. Miller, M.K. Shivakumar, S.L. Risacher, A.J. Saykin, S. Lee, K. Nho, D. Kim, Codon bias among synonymous rare variants is associated with Alzheimer's disease imaging biomarker, Pac. Symp. Biocomput. 23 (2018) 365–376.
[41] I.B. Rogozin, E.M. Gertz, P.V. Baranov, E. Poliakov, A.A. Schaffer, Genome-wide changes in protein translation efficiency are associated with autism, Genome Biol. Evol. 10 (2018) 1902–1919.
[42] C.E. Brule, E.J. Grayhack, Synonymous codons: choose wisely for expression, Trends Genet. 33 (2017) 283–297.
[43] R.J. Weatheritt, M.M. Babu, Evolution. The hidden codes that shape protein evolution, Science 342 (2013) 1325–1326.
[44] F. Wright, The 'effective number of codons' used in a gene, Gene 87 (1990) 23–29.
[45] F. Tulloch, N.J. Atkinson, D.J. Evans, M.D. Ryan, P. Simmonds, RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies, Elife 3 (2014), e04531.
[46] S.A. Babayan, R.J. Orton, D.G. Streicker, Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes, Science 362 (2018) 577–580.
[47] S.H. Shen, C.B. Stauft, O. Gorbatsevych, Y. Song, C.B. Ward, A. Yurovsky, S. Mueller, B. Futcher, E. Wimmer, Large-scale recoding of an arbovirus genome to rebalance its insect versus mammalian preference, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 4749–4754.
[48] B. Futcher, O. Gorbatsevych, S.H. Shen, C.B. Stauft, Y. Song, B. Wang, J. Leatherwood, J. Gardin, A. Yurovsky, S. Mueller, E. Wimmer, Reply to Simmonds et al.: Codon pair and dinucleotide bias have not been functionally distinguished, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) E3635–E3636.
[49] V. Simonyan, K. Chumakov, H. Dingerdissen, W. Faison, S. Goldweber, A. Golikov, N. Gulzar, K. Karagiannis, P. Vinh Nguyen Lam, T. Maudru, O. Muravitskaja, E. Osipova, Y. Pan, A. Pschenichnov, A. Rostovtsev, L. Santana-Quintero, K. Smith, E.E. Thompson, V. Tkachenko, J. Torcivia-Rodriguez, A. Voskanian, Q. Wan, J. Wang, T.J. Wu, C. Wilson, R. Mazumder, High-Performance Integrated Virtual Environment (HIVE): a robust infrastructure for next-generation sequence data analysis, Database (Oxford) 2016 (2016).
[50] 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.