Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

The Kazusa codon usage database, CoCoPUTs, and the value of up-to-date codon usage statistics

ARTICLE INFO

Keywords:
Codon usage
Genbank
Dengue virus
Kazusa codon usage database
Codon/codon pair usage tables
HIVE codon usage table database

To the Editor,

In 2017, we established the HIVE Codon Usage Table Database (HIVE-CUTs) (Athey et al., 2017), a database of codon usage for all species with coding sequences in either Genbank, the NCBI sequence database, or Refseq, the NCBI Reference Sequence Database. HIVE-CUTs has recently been superseded by the Codon/Codon Pair Usage Table Data (CoCoPUTs) (Alexaki et al., 2019), which contains not only codon usage data, but also data for codon pair and dinucleotide usage.

In the research paper "Codon adaptation biases among sylvatic and urban genotypes of Dengue" (Cunha et al., 2018), Cuhna MDP et al. analyzed the codon adaptation of DENV serotype 2 genotypes from urban and sylvatic habitats in comparison to human and Ae. aegypti hosts. The authors conclude that DENV-2 genotypes have a higher Codon Adaptation Index (CAI) (Sharp and Li, 1987) to humans that to Ae. aegypti; however, their analysis relies on the Kazusa Codon Usage Database (Nakamura et al., 2000), which has not been updated since 2007.

The Kazusa Codon Usage Database sources its data from NCBI Genbank release 160, June 15, 2007, therefore it contains data parsed from 3,027,973 complete protein coding sequences (CDS's) for 35,799 species/organelle pairs. The Kazusa Codon Usage database was groundbreaking at the time of its publication and has been widely used in analyses and tools involving codon usage but is now increasingly becoming obsolete.

To note, CoCoPUTs is updated quarterly following each new Genbank release, and draws from Refseq, the NCBI Reference Sequence Database. For this reason, CoCoPUTs analyzes 725,494,506 CDS's for 1,431,241 species/organelle pairs. CoCoPUTs contains data for more species/organelles, more CDS's for each species/organelle, and allows a user to agglomerate codon usage data for higher taxa.

Repeating the analysis with updated data from CoCoPUTs indicates that CAI values (as shown in Fig. 1 of Cuhna MDP et al.) were substantially underestimated for mosquito, while they were less affected for *Homo sapiens* (Fig. 1). This is unsurprising, because the authors didn't use Kazusa as a data source for *Homo sapiens* codon usage.

Using a different dataset may result in some large disparities in codon usage for some species, particularly for species that have been more extensively studied in recent years. The median Euclidean distance between CoCoPUTs and Kazusa codon profiles over all species/organelle pairs contained in both is 33.80. This disparity exists because CoCoPUTs has more CDS's, more codons, fewer zero entries, and more entropy, a measure of information contained in a probability distribution, for species/organelles contained in both databases (Table 1). In total, 61.34% of species/organelle pairs in both databases have lower entropy in the Kazusa database, and 55.76% have more zeroes in their codon profile in Kazusa.

Many of the species with the biggest distances between CoCoPUTs and Kazusa are marsupials (Supplementary Table S1). Many of the Kazusa entries for these species contain one CDS (Sperm protamine P1, PRM1), with many zeroes in the codon usage table. The small number of codons for these entries profoundly impacts the codon distribution. CoCoPUTs has more CDS's and codons, less zero entries, and higher entropy for these species.

As an alternative measure of similarity between codon profiles, we also examined correlations between Kazusa and CoCoPUTs codon profiles (Supplementary Table S2). The median correlation between Kazusa and CoCoPUTs codon profiles is 0.956, and the mean correlation is 0.901. The species/organelles with the lowest correlations tend to be bats (genomic) and birds (mitochondrial). Both types have only one CDS in the Kazusa database: NADH-ubiquinone oxidoreductase chain 6 (MT-ND6) in the mitochondrial bird data, Sperm Protamine 1 (PRM1) in the genomic bat data.

Although the historic contribution of the Kazusa database is undeniable, it contains incomplete data from an outdated Genbank release. We therefore encourage scientists in the future to use CoCoPUTs, https://hive.biochemistry.gwu.edu/review/codon2, which is regularly updated, and, as shown in the analysis provided, contains more accurate and informative data. As we demonstrated, incompleteness of data in the Kazusa resource will affect the overall conclusions of studies relying on codon usage statistics.

1. Methods

Kazusa Codon Usage Database data was downloaded from the Kazusa FTP server. Codon counts were summed for all CDS's for each

https://doi.org/10.1016/j.meegid.2019.05.010
Received 6 May 2019; Received in revised form 10 May 2019; Accepted 13 May 2019
Available online 15 May 2019
1567-1348/ Published by Elsevier B.V.

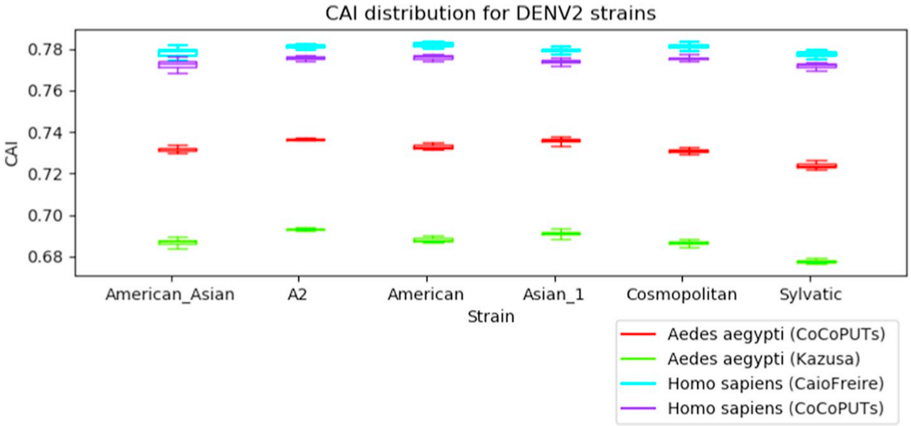

Fig. 1. CAI distributions of Dengue Virus 2 genomes, using *Aedes aegypti* and *Homo sapiens* CUTs from CoCoPUTs, Kazusa Codon Usage Database, and Caio Freire's Github.

Table 1
Comparison of Kazusa and CoCoPUTs data. Mean and median values of the number of CDS's, codons, zero entries, and entropy are computed over all species/organelle pairs in Kazusa, CoCoPUTs, both CoCoPUTs and Kazusa, and CoCoPUTs but not in Kazusa.

	Number of species/organelle pairs	Number of coding sequences		Number of codons		Number of zero entries		Entropy	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
All Kazusa entries	35,792	2	84.33	551	29,255.1	4	7.12	3.77	3.73
All CoCoPUTs entries	1,307,999	1	502.2	310	159,195	16	15.36	3.54	3.54
CoCoPUTs entries also in Kazusa	34,300	10	6996	2747	2,277,833	2	3.4	3.82	3.79
CoCoPUTs entries not in Kazusa	1,273,699	1	327.3	287	102,141	17	15.68	3.53	3.53

taxid/organelle pair.

Codon usage data was downloaded from the CoCoPUTs server, backend version “All RefSeq and GenBank March 2019”. Refseq and Genbank codon counts were added for each taxid/organelle pair, because CoCoPUTs only compiles Genbank data for assemblies not found in Refseq.

Data was compiled and analyzed with Python 3.5.5, using Biopython, SciPy, Matplotlib, and in-house scripts, which are included in supplementary materials.

Codon usage is measured as codon frequency (per 1000 codons), $F(c) = 1000(N_c/\sum_i N_i)$, where N_c is the count for codon c.

A codon profile or codon usage profile for a species/organelle is an array of codon frequencies (per 1000 codons) for each codon.

Distance between codon profiles is Euclidean distance between Kazusa and CoCoPUTs codon usage profiles, using codon frequencies per 1000 codons, $\sqrt{\sum_c (F_{Kazusa}(c) - F_{CoCoPUTs}(c))^2}$, where $F_{Kazusa}(c)$ and $F_{CoCoPUTs}(c)$ are the codon frequencies for codon c from Kazusa and CoCoPUTs, respectively.

Shannon entropy is a measure of the information contained in a probability distribution. For codon usage data, entropy is computed as $\sum_c - p_c \ln p_c$, where $p_c = N_c/\sum_i N_i = F(c)/\sum_i F(i.)$

The number of zero entries in a codon profile is self-explanatory, $\sum_c \begin{cases} 1 \text{ if } F(c) = 0 \\ 0 \text{ else} \end{cases}$.

CAI is defined as in Sharp and Li (Sharp and Li, 1987), and calculated on all DENV2 CDS's listed in Supplementary Table S1 of Cuhna MDP et al., using codon usage tables from CoCoPUTs, Kazusa Codon Usage Database, and the Github of Caio César de Melo Freire.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2019.05.010.

Acknowledgements

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.

The authors have declared that no competing interests exist.

References

Alexaki, A., Kames, J., Holcomb, D.D., Athey, J., Santana-Quintero, L.V., Lam, P.V., ... Kimchi-Sarfaty, C., 2019. Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant Gene Design. J. Mol. Biol. https://doi.org/10.1016/j.jmb.2019.04.021.

Athey, J., Alexaki, A., Osipova, E., Rostovtsev, A., Santana-Quintero, L.V., Katneni, U., ... Kimchi-Sarfaty, C., 2017. A new and updated resource for codon usage tables. BMC

Bioinforma. 18. https://doi.org/10.1186/s12859-017-1793-7.

Cunha, M.D., Ortiz-Baez, A.S., Freire, C.C., Zanotto, P.M., 2018. Codon adaptation biases among sylvatic and urban genotypes of dengue virus type 2. Infect. Genet. Evol. 207–211. https://doi.org/10.1016/j.meegid.2018.05.017.

Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 292.

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

David D. Holcomb, Aikaterini Alexaki, Upendra Katneni,
Chava Kimchi-Sarfaty[*]
Division of Plasma Protein Therapeutics, Office of Tissues and Advanced Therapies, Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, USA
E-mail address: Chava.kimchi-sarfaty@fda.hhs.gov (C. Kimchi-Sarfaty).

[*] Corresponding author.