SOFTWARE TOOL ARTICLE

vhcub: Virus-host codon usage co-adaptation analysis [version 1; peer review: 2 approved]

Ali Mostafa Anwar[1], Mohamed Soudy[2], Radwa Mohamed[1]

[1]Department of Genetics, Faculty of Agriculture, Cairo University, Cairo, 12613, Egypt
[2]Bioinformatics program, Faculty of Computer Science, Ain Shams University, Ain Shams, Egypt

v1 **First published:** 23 Dec 2019, **8**:2137 (https://doi.org/10.12688/f1000research.21763.1)
Latest published: 23 Dec 2019, **8**:2137 (https://doi.org/10.12688/f1000research.21763.1)

Abstract

Viruses show noticeable evolution to adapt and reproduce within their hosts. Theoretically, patterns and factors that affect the codon usage of viruses should reflect evolutionary changes that allow them to optimize their codon usage to their hosts. Some software tools can analyze the codon usage of organisms; however, their performance has room for improvement, as these tools do not focus on examining the codon usage co-adaptation between viruses and their hosts. This paper describes the *vhcub* R package, which is a crucial tool used to analyze the co-adaptation of codon usage between a virus and its host, with several implementations of indices and plots. The tool is available from: https://cran.r-project.org/web/packages/vhcub/.

Keywords

Evolution, Natural selection, Adaptation, Viruses, Codon Usage Bias, R, RStudio

This article is included in the RPackage gateway.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 1 23 Dec 2019	✓ report	✓ report

1 **Raj Kumar Singh**, ICAR-Indian Veterinary Research Institute [Deemed University], Izatnagar, India

2 **Oscar Leonardo Ramírez Suárez**, Ginic-Hus Group Universidad ECCI, Bogotá, Colombia
Adriana Patricia Corredor-Figueroa, Universidad ECCI, Bogota, Colombia

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Ali Mostafa Anwar (ali.mo.anwar@std.agr.cu.edu.eg)

Author roles: Anwar AM: Conceptualization, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Soudy M**: Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Mohamed R**: Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Anwar AM, Soudy M and Mohamed R. ***vhcub*: Virus-host codon usage co-adaptation analysis [version 1; peer review: 2 approved]** F1000Research 2019, **8**:2137 (https://doi.org/10.12688/f1000research.21763.1)

First published: 23 Dec 2019, **8**:2137 (https://doi.org/10.12688/f1000research.21763.1)

Introduction

During the translation process from mRNAs to proteins, information is transmitted in the form of triple nucleotides, named codons, which encode amino acids. Multiple codons that encode one amino acid are known as synonymous codons. Studies concerning different organisms report that synonymous codons are not used uniformly within and between genes of one genome, a phenomenon known as codon usage bias (CUB)[1,2]. Since viruses rely on the tRNA pool of their hosts in the translation process, previous studies suggest that translation selection or/and directional mutational pressure act on the codon usage of the viral genome to optimize or deoptimize it towards the codon usage of their hosts[3,4].

Tools and packages are available to analyze codon usage, e.g. coRdon[5], but there is no package available that focuses on the examination of codon usage co-adaptation between viruses and their hosts. vhcub is a package implemented in R, which aims to easily analyze the co-adaptation of codon usage between a virus and its host. vhcub measures several codon usage bias measurements, such as effective number of codons (ENc)[6], codon adaptation index (CAI)[7], relative codon deoptimization index (RCDI)[8], similarity index (SiD)[9], synonymous codon usage orderliness (SCUO)[10], and relative synonymous codon usage (RSCU)[10]. It also provides a statistical dinucleotide over- and under-representation with three different models.

Methods

Implementation

vhcub imports Biostrings[11], seqinr[12] and stringr[13] to handle fasta files and manipulate DNA sequences. Also, it imports coRdon[5], which is used to estimate different CUB measures.

vhcub first converts the fasta format to data.frame type, to efficiently maintain and calculate different indices implemented in the package. Table 1 describes all the functions available in vhcub, and the result returned from each. Also, it contains references to the equations used to estimate each measure. Furthermore, vhcub uses ggplot2[14] to

Table 1. Functions available in vhcub, and the result returned from each one.

Function name	Description	Value
fasta.read	Read fasta formate and convert it to data frame	A list with two data.frames; the first one for virus DNA sequences and the second one for the host.
CAI.values	Measure the Codon Adaptation Index (CAI) using Sharp and Li (1987)[7] equation, of DNA sequence.	A data.frame containing the computed CAI values for each DNA sequences within df.fasta.
dinuc.base	A measure of statistical dinucleotide over- and under-representation; by allows for random sequence generation by shuffling (with/without replacement) of all bases in the sequence[13].	A data.frame containing the computed statistic for each dinucleotide in all DNA sequences within df.virus.
dinuc.codon	A measure of statistical dinucleotide over- and underrepresentation; by allows for random sequence generation by shuffling (with/without replacement) of codons[13].	A data.frame containing the computed statistic for each dinucleotide in all DNA sequences within df.virus.
dinuc.syncodon	A measure of statistical dinucleotide over- and underrepresentation; by allows for random sequence generation by shuffling (with/without replacement) of synonymous codons[13].	A data.frame containing the computed statistic for each dinucleotide in all DNA sequences within df.virus.
ENc.values	Measure the Effective Number of Codons (ENc) of DNA sequence. Using its modified version (Novembre, 2002)[6].	A data.frame containing the computed ENc values for each DNA sequences within df.fasta.
GC.content	Calculates overall GC content as well as GC at first, second, and third codon positions.	A data.frame with overall GC content as well as GC at first, second, and third codon positions of all DNA sequence from df.virus.
RCDI.values	Measure the Relative Codon Deoptimization Index (RCDI)[8] of DNA sequence.	A data.frame containing the computed ENc values for each DNA sequences within df.fasta.
RSCU.values	Measure the Relative Synonymous Codon Usage (RSCU)[7] of DNA sequence.	A data.frame containing the computed RSCU values for each codon for each DNA sequences within df.fasta.

Function name	Description	Value
SCUO.values	Measure the Synonymous Codon Usage Eorderliness (SCUO) of DNA sequence using Wan *et al.*, 2004[10] equation.	A data.frame containing the computed SCUO values for each DNA sequences within df.fasta.
SiD.value	Measure the Similarity Index (SiD) between a virus and its host codon usage[15].	A numeric represent a SiD value.
PR2.plot	Make a Parity rule 2 (PR2) plot[16], where the AT-bias [A3/(A3 +T3)] at the third codon position of the four-codon amino acids of entire genes are the ordinate and the GC-bias [G3/(G3 +C3)] is the abscissa. The centre of the plot, where both coordinates are 0.5, is where A = U and G = C (PR2), with no bias between the influence of the mutation and selection rates.	A ggplot object.
ENc.GC3plot	Make an ENc-GC3 scatterplot[17]. Where the y-axis represents the ENc values and the x-axis represents the GC3 content. The red fitting line shows the expected ENc values when codon usage bias affected solely by GC3.	A ggplot object.

visualize two important plots named ENc-GC3 plot (Figure 2) and PR2-plot (Figure 3), which help to explain the factors influencing a virus's evolution concerning its CUB.

Operation

vhcub was developed using R and is available on CRAN. It is compatible with Windows, and major Linux operating systems. The package can be installed as:

```
install.packages( "vhcub" )
```

Figure 1 describes the vhcub workflow. It starts with reading the fasta files for a virus and its host. After, nucleotide content analysis, codon usage bias analysis on genes and codon level (marked by the red boxes in Figure 1) can be applied independently (the blue boxes in Figure 1). However, within the same analysis, some measures rely on others. For example, the reference set of genes used to estimate a virus codon adaptation index was defined based on the effective number of codons of its host. Finally, the orange boxes in Figure 1 represent the two plots (ENc-GC3 plot and PR2-plot).

Use cases

Using vhcub to study the CUB of a virus, its host and the co-adaptation between them is straightforward. As an example, we have used the coding sequences for *Escherichia virus T4* and its host *Escherichia coli* in the form of fasta format.

```
# First to call the library
library("vhcub")

# To read both files at the same time as a data.frame
# Using fasta.read() function
# virus.fasta = directory path to the virus fasta file
# host.fasta = directory path to the host fasta file.

fasta <- fasta.read (virus.fasta = "EscherichiavirusT4.fasta",
                     host.fasta = "Escherichiacoli.fasta")

fasta.T4 <- fasta[[1]]
fasta.Ecoli <- fasta[[2]]
```

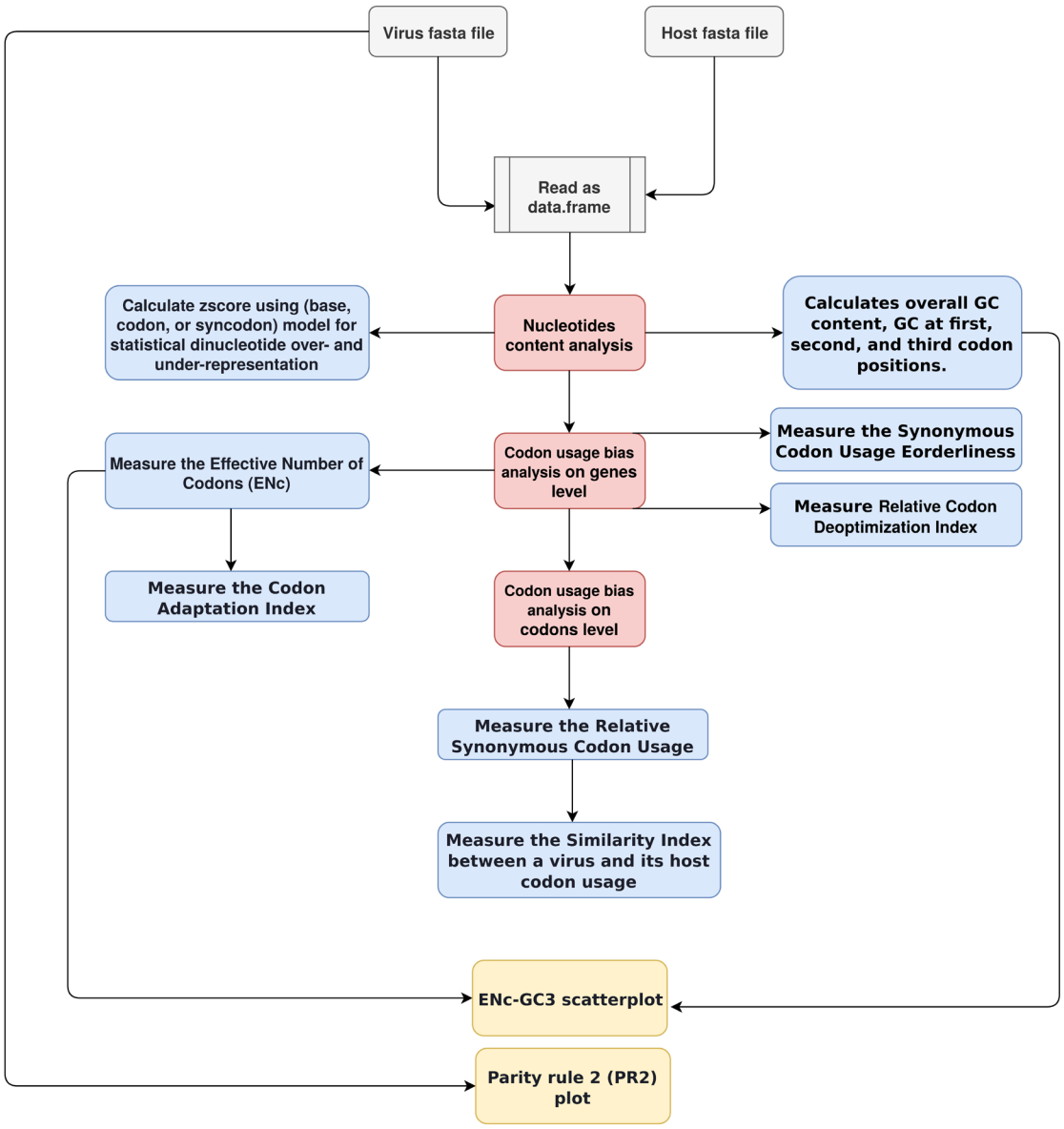

Figure 1. vhcub workflow, to analyze virus-host codon usage co-adaptation. The white boxes represent the input fasta files. The red boxes represent three main analysis, each with different measures (the blue boxes), and the orange boxes represent ENc-GC3 plot and PR2-plot.

As mentioned before, each category of analysis could be applied independently. Hence, this example will show only the codon usage bias analysis at the codon level.

```
# To estimate the similarity index (SiD) between E.coli T4 virus and E.coli

#First Calculate the Relative Synonymous Codon Usage (RSCU) for both of them
rscu.T4 <- RSCU.values(fasta.T4)
rscu.Ecoli <- RSCU.values(fasta.Ecoli)

# Then, the SiD could be calculated as
SiD <- SiD.value(rscu.Ecoli, rscu.T4)
```

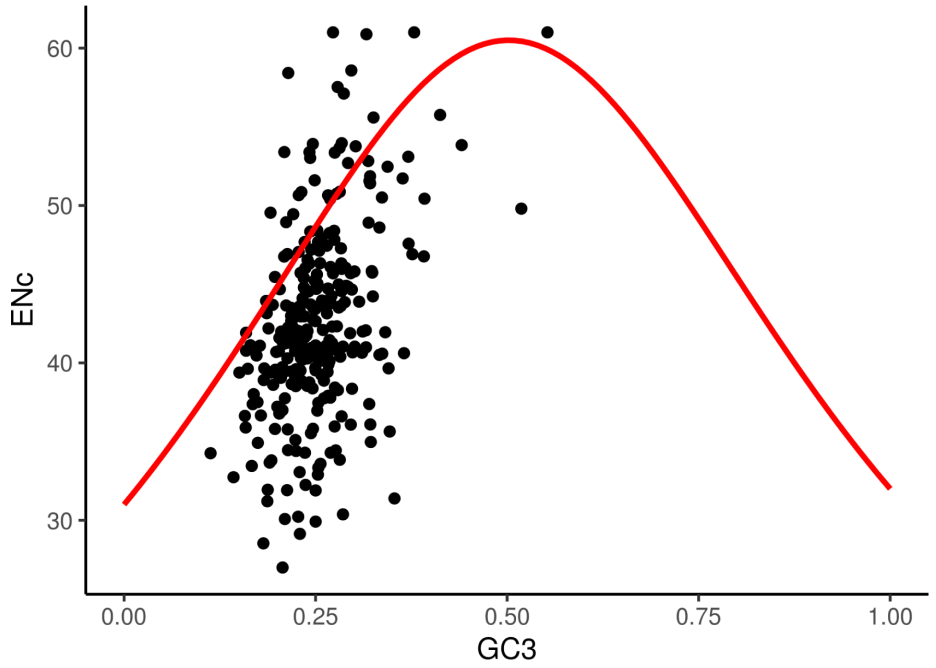

Figure 2. ENc-GC3 plot showing the values of the ENc versus the GC3 content for the virus (Escherichia virus T4) CDS, the solid red line represents the expected ENc values if the codon bias is affected by GC3s only.

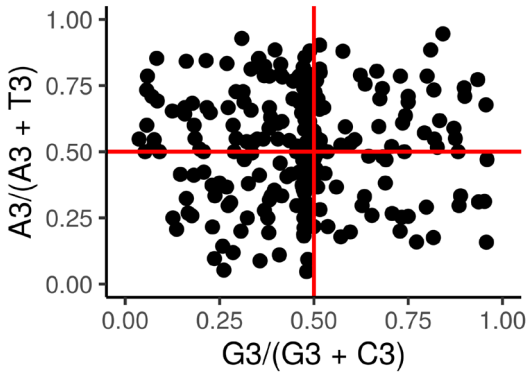

Figure 3. PR2-plot showing CDS of the virus (Escherichia virus T4), plotted based on their GC bias [G3/(G3 + C3)] and AT bias [A3/(A3 + T3)] in the third codon position, the two solid red lines represent both coordinates (ordinate and abscissa) equal to 0.5, where A = T and G = C.

SiD measures the effect of the codon usage bias of the *E. coli* on *E. coli* T4 virus. In general, SiD ranged from 0 to 1 with higher values indicating that the host has a dominant effect on the usage of codons. In this example, SiD is approximately equal to 0.491. Which means that E. *coli* does not dominate *E. coli* T4 CUB. Also, this code generates RSCU values for each codon in each gene from both organisms and can be used for further analysis.

Conclusions

vhcub depends only on DNA sequences as input and can compute different measures of CUB for viruses, such as ENc, CAI, SCUO, and RCDI (Table 1). It can also be used to study the association between viruses and their hosts' RSCU and SiD. There are many possible directions for future work; further versions will execute more indices, plots, and statistical analysis, to facilitate the workflow for examining the adaptations of viruses' CUB in the R environment.

Data availability

Escherichia virus T4 fasta file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/836/945/GCF_000836945.1_ViralProj14044

Escherichia coli fasta file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_cds_from_genomic.fna.gz

Software availability

Software available from: https://CRAN.R-project.org/package=vhcub

Source code available from: https://github.com/AliYoussef96/vhcub

Archived source code as at time of publication: http://doi.org/10.5281/zenodo.3572391[18]

License: GPL-3

References

1. Behura SK, Severson DW: **Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes.** *PLoS One.* 2012; **7**(8): e43111.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Boël G, Letso R, Neely H, *et al.*: **Codon influence on protein expression in E. coli correlates with mRNA levels.** *Nature.* 2016; **529**(7586): 358–363.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Burns CC, Shaw J, Campagnoli R, *et al.*: **Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region.** *J Virol.* 2006; **80**(7): 3259–3272.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Cladel NM, Hu J, Balogh KK, *et al.*: **CRPV genomes with synonymous codon optimizations in the CRPV E7 gene show phenotypic differences in growth and altered immunity upon E7 vaccination.** *PLoS One.* 2008; **3**(8): e2947.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Elek A, Kuzman M, Vlahovicek K: **coRdon: Codon Usage Analysis and Prediction of Gene Expressivity.** R package version 1.0.3. 2019.
Reference Source
6. Novembre JA: **Accounting for background nucleotide composition when measuring codon usage bias.** *Mol Biol Evol.* 2002; **19**(8): 1390–1394.
PubMed Abstract | Publisher Full Text
7. 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**(3): 1281–1295.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Puigbò P, Aragonès L, Garcia-Vallvé S: **RCDI/eRCDI: a web-server to estimate codon usage deoptimization.** *BMC Res Notes.* 2010; **3**(1): 87.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Zhou JH, Zhang J, Sun DJ, *et al.*: **The distribution of synonymous codon choice in the translation initiation region of dengue virus.** *PLoS One.* 2013; **8**(10): e77239.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Wan XF, Xu D, Kleinhofs A, *et al.*: **Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes.** *BMC Evol Biol.* 2004; **4**(1): 19.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Pagès H, Aboyoun P, Gentleman R, *et al.*: **Biostrings: Efficient manipulation of biological strings.** R package version 2.50.2. 2019.
Reference Source
12. Charif D, Lobry JR: **SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis.** In: U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, editors. *Structural approaches to sequence evolution: Molecules, networks, populations, Biological and Medical Physics, Biomedical Engineering.* Springer Verlag, New York. 2007; 207–232.
Publisher Full Text
13. Wickham H: **stringr: Simple, Consistent Wrappers for Common String Operations.** R package version 1.4.0. 2019.
Reference Source
14. Wickham H: **ggplot2: Elegant Graphics for Data Analysis.** Springer-Verlag New York. 2016.
Reference Source
15. He Z, Gan H, Liang X: **Analysis of Synonymous Codon Usage Bias in Potato Virus M and Its Adaption to Hosts.** In. *Viruses.* 2019; **11**(8). pii: E752.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Xiang H, Zhang R, Butler RR 3rd, *et al.*: **Comparative Analysis of Codon Usage Bias Patterns in Microsporidian Genomes.** *PLoS One.* 2015; **10**(6): e0129223.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Butt AM, Nasrullah I, Qamar R, *et al.*: **Evolution of codon usage in zika virus genomes is host and vector specific.** *Emerg Microbes Infect.* 2016; **5**(10): e107.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Youssef A: **AliYoussef96/vhcub: Virus-Host Codon Usage Co-Adaptation Analysis (Version v1.0.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3572391

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 27 March 2020

https://doi.org/10.5256/f1000research.23991.r61560

✔ **Oscar Leonardo Ramírez Suárez**
Ginic-Hus Group Universidad ECCI, Bogotá, Colombia
Adriana Patricia Corredor-Figueroa
Universidad ECCI, Bogota, Colombia

- From the technical point of view, the vhcub R package looks quite reliable and well supported. However, there is only one example illustrating it. Moreover, this example shows the mean value for SiD in its range (i.e., 0.491 or approx. 0.5), which is great but makes us wondering if this package could give expected values for other cases. If that is possible, could the authors include a couple of examples where the SiD value were below 0.5 and above 0.5?

- The tool is very interesting and captivating. The advantages that R offers are infinite, so I consider that it would be invaluable to exploit the output that R offers. It is clear in the article that the algorithm only allows the entry of DNA sequences in Fasta format, although there are other very simple tools to use to transcribe from RNA to DNA, or from RNA- to RNA + and DNA, it would be very nice to use the same R package to carry out this step, especially considering that from biological tests we not only analyze one sequence but many. To the same extent, I consider that the figures presented in the article should be more discussed from a biological point of view, they could be more informative.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Virology, molecular biology, infectious diseases

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 11 February 2020

https://doi.org/10.5256/f1000research.23991.r58975

Raj Kumar Singh
Facility for Research and Training on Bioassays and Biosensor, Division of Veterinary Biotechnology, ICAR-Indian Veterinary Research Institute [Deemed University], Izatnagar, Uttar Pradesh, India

Viruses in the course of their evolution would optimize their codon usage to their hosts. They rely on the tRNA pool of their hosts in the translation process. Though tools for analyzing the codon usage of organisms are available, none of them focus on examining the codon usage co-adaptation between viruses and their hosts. This software, *vhcub,* is a tool used to analyze the co-adaptation of codon usage between a virus and its host. This may also help to predict the possible mutations that would accumulate in the virus vis - a - vis its host(s), thereby showing the readiness for the control and prevention of the disease.

General comments

1. Corrections in the text
   - Spelling of formate may be corrected to format in the second column X first row of Table 1
   - Please define df.fasta
   - In Third column X sixth row and Third column X eight row are one and the same - please explain or correct

Specific
Whether it can be used in Eukayotes?

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Viral immunology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 11 Feb 2020

Ali Mostafa, Department of Genetics, Faculty of Agriculture, Cairo University, Egypt

General comments (Corrections in the text)

- Comment: The spelling of formate may be corrected to format in the second column X first row of Table 1.
  Response: I will make this correction during the article revisions.
- Comment: Please define df.fasta
  Response: (df.host) as well as (df.virus) are just variables names for data frames holds host genes and virus genes, respectively. The definition will be added during the article revisions.
- Comment: In Third column X sixth row and Third column X eight row are one and the same - please explain or correct
  Response: In the third column X eight row. It will be corrected from ENc to RCDI.

Specific comments

- Comment: Whether it can be used in Eukaryotes?
  Response: The translation codon table (The Genetic Codes Tables) number could be changed to any table number (As defined by NCBI https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi), in vhcub. For example in the function named (CAI.values()), one can pass an argument genetic.code="11" for bacterial codon table or genetic.code="1" for eukaryotes. Hence, yes, the host can be Prockayotic or Eukaryotic (vhcub can be used for Eukaryotes).

Competing Interests: No competing interests were disclosed.

Reviewer Response 13 Feb 2020

Raj Kumar Singh, ICAR-Indian Veterinary Research Institute [Deemed University], Izatnagar, India

The authors have accepted to do the necessary changes in the revised version and as well they have answered to my query.

Competing Interests: No competing Interests