The Journal of Open Source Software

Resvidex: An R package for HRSV classification

Marco Cacciabue [1,2¶], **Nahuel Axel Fenoglio** [3], **Melina Obregón** [3], **Stephanie Goya** [4], **and Mariana Viegas** [5,1]

1 Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina. **2** Departamento de Ciencias Básicas, Universidad Nacional de Luján, Luján, Argentina. **3** Universidad Nacional de Luján, Luján, Argentina. **4** Department of Laboratory Medicine and Pathology, University of Washington Medical Center, Seattle, WA, USA **5** Laboratorio de Salud Pública, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Buenos Aires, Argentina ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

The human respiratory syncytial virus (HRSV) is one of the leading causes of acute lower respiratory tract infection in children, elderly and immunocompromised individuals. Below species level, there are two antigenic groups: HRSV subgroup A (HRSV-A) and B (HRSV-B). Within each subgroup, genotypes are defined based on statistically supported phylogenetic clades that can be inferred with the second hypervariable region (2HR) of the G gene, which encodes the attachment glycoprotein and exhibits the highest genetic and antigenic variability.

Advanced machine learning techniques have proven to make accurate predictions, using algorithms that reveal patterns in large datasets. In the analysis of viral data, machine learning methods have been recently implemented, for example, in: COVIDEX, a tool that classifies complete genome nucleotide sequences of SARS-CoV-2 into lineages (Cacciabue et al., 2022), or iNFINITY, another tool used to make human influenza virus classification into subtypes and clades (Cacciabue & Marcone, 2023).

Statement of need

`Resvidex` is a tool based on alignment-free machine learning for HRSV classification into subtypes and clades. Resvidex is a web application that runs on an internet connection without any installation and has a user-friendly interface. It is fast, sensitive, specific, and ready to implement. Additionally, it is available to run locally for R and Rstudio users as an R package.

The overall classification algorithm that `Resvidex` uses is divided into three phases:

1. The first phase loads the user data in a multifasta format and performs the k-mer counting operation using the k-mer package (Wilkinson, 2018). Each k-mer count is normalized over the k-mer size ($k = 6$) and the sequence length.
2. The second phase calls the ranger package (Wright & Ziegler, 2015) predict function using a pre-trained random forest model and obtains a probability score based on the rule of majority vote. From this, the app obtains the score for each query sequence classification, the proportion of N bases in the genome, and the genome length.
3. Finally, two tables are created, one showing the sequences that passed all the quality checks and another with sequences that did not pass some of the filter steps. These filters controls: that each sequence obtained a probability score of 0.4 or more, that the sequence length is close to the expected sequence length for the classification model for a factor of no more that 50%, and that the percentage of ambiguous bases in the sequence (N) is not larger than 2%. A brief report can be produced including the results table, date of analysis, and model information.

`Resvidex` was designed to be used by researchers who want to classify their samples of HRSV. It is used at the moment as a part of a HRSV phylogeny investigation (Goya et al., 2024)

Acknowledgements

References

Cacciabue, M., Aguilera, P., Gismondi, M. I., & Taboga, O. (2022). Covidex: An ultrafast and accurate tool for SARS-CoV-2 subtyping. *Infection, Genetics and Evolution*, *99*, 105261. https://doi.org/10.1016/j.meegid.2022.105261

Cacciabue, M., & Marcone, D. N. (2023). INFINITy: A fast machine learning-based application for human influenza A and B virus subtyping. *Influenza and Other Respiratory Viruses*, *17*(1), e13096. https://doi.org/10.1111/irv.13096

Goya, S., Ruis, C., Neher, R. A., Meijer, A., Aziz, A., Hinrichs, A. S., Von Gottberg, A., Roemer, C., Amoako, D. G., Acuña, D., McBroome, J., Otieno, J. R., Bhiman, J. N., Everatt, J., Muñoz-Escalante, J. C., Ramaekers, K., Duggan, K., Presser, L. D., Urbanska, L., … Viegas, M. (2024). *The unified proposal for classification of human respiratory syncytial virus below the subgroup level* [Preprint]. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2024.02.13.24302237

Wilkinson, S. (2018). Kmer: An r package for fast alignment-free clustering of biological sequences. In *GitHub repository*. https://doi.org/10.5281/zenodo.1227690

Wright, M. N., & Ziegler, A. (2015). *Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R*. https://doi.org/10.48550/ARXIV.1508.04409