Received: 22 September 2022 | Revised: 3 January 2023 | Accepted: 8 January 2023

DOI: 10.1111/irv.13096

LETTER TO THE EDITOR

WILEY

INFINITy: A fast machine learning-based application for human influenza A and B virus subtyping

Influenza viruses are one of the main agents causing acute respiratory infections (ARI) in humans resulting in a large amount of illness and death globally.[1,2] The influenza viruses classification is based on the nomenclature proposed by the World Health Organization (WHO)[3] that is widely accepted and used by the medical and scientific communities throughout the world. Since the pandemic in 2009, two subtypes of human influenza A viruses, A(H1N1)pdm09 and A(H3N2), and two lineages of influenza B, B/Victoria and B/Yamagata, have been responsible for the vast majority of cases each year. Within each subtype and lineage, different clades and genetic groups were described to reflect the continuous viral evolution, driven by antigenic drift. The WHO Global Influenza Surveillance and Response System (GISRS) studies human influenza viruses from >110 countries, to monitor circulating strains, understand epidemiology and evolution, and contribute to verify the vaccine effectiveness and update its formulation each year.[4,5] A growing number of laboratories and research centers is contributing to this initiative by sequencing the whole viral genome or the hemagglutinin (HA) gene from local strains.

Influenza clade classification is usually performed by phylogenetic analysis of HA gene sequences from circulating strains along with reference sequences, which is a time-consuming process and requires specific training and equipment. Alternatively, this can be done by comparing amino acid substitutions, either manually or by using in-house scripts. While there are currently specific tools available for influenza classification,[6–8] they have several limitations such as: (a) they require an alignment of the input data against reference sequences (which can be computationally expensive), (b) requirement of multiple ad hoc programs installed, (c) users should be familiar with the command line, (d) users must create a template containing clade-defining amino acid pattern by position, (e) only classifies sequences into type A or B and subtype/lineage but cannot discern clades or genetic groups, and (f) take into account only the most prevalent and recent influenza clades.

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,[9] a recent application for avian influenza clade classification,[10] the prediction of phenotypes for human influenza A from proteomic input,[11] and detection of new variants using ensemble learning.[12]

In this sense, we developed INFINITy, a tool based on alignment-free machine learning for human influenza virus classification into subtypes and clades. INFINITy 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. Furthermore, two docker images are available to secure the reproducibility of the results.

INFINITy includes two classification models: one for complete HA sequences (FULL HA, for whole gene sequence length of 1700 bp) and other for the HA1 subunit coding sequence (HA1, for the initial 1030 bp of the HA gene). The influenza classification comprises 75 clades or genetic groups: 25 for A(H1N1)pdm09, 32 for A(H3N2), and 14 for B/Victoria and 4 for B/Yamagata (supporting information Table S1).

The overall classification algorithm 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.[13] 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[14] predict function using one of the two pre-trained random forest models (FULL HA or HA1) 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 (FULL HA 1700 or HA1 1030) 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 (Figure 1).

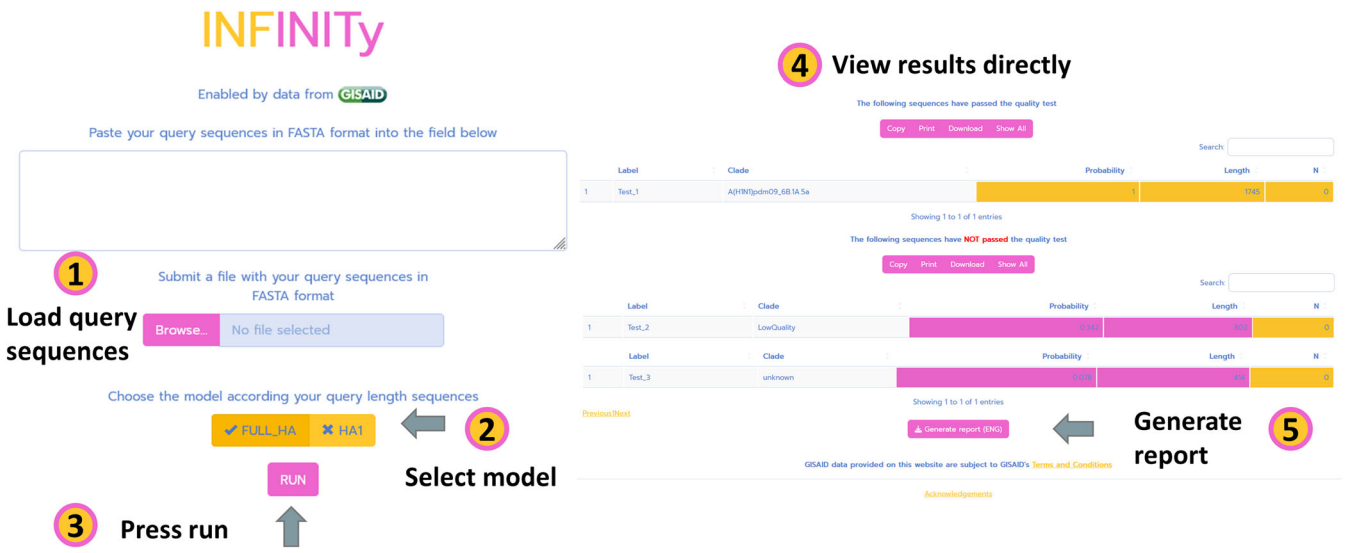

FIGURE 1 Overview of the INFINITy application. The user loads a sequence file, or copy and paste the sequences, selects the corresponding model, and presses RUN. Two results tables will be shown, showing the sequences that passed the quality controls and those that did not. Although all sequences are classified, the user should carefully interpret the results considering the quality control for each one. Sequences that did NOT passed the quality filters are shown as "LowQuality", and those sequences with a probability score below a value of 0.2 are shown as "unknown". Finally, the user can download an automatic report.

In order to train the classification models, a reference dataset was created by downloading complete HA sequences of influenza A(H1N1)pdm09, A(H3N2), B/Victoria, and B/Yamagata from GISAID. We defined the influenza clade and subclade for each sequence by analyzing their amino acid composition and the combination of signature position mutations, according to the WHO nomenclature. A phylogenetic analysis by influenza type and lineage was used to confirm the classification. The final HA gene dataset includes a total of 11,316 sequences with an average length of 1700 bp: 2957 influenza A(H1N1)pdm09, 3112 A(H3N2), 2963 B/Victoria, and 2284 B/Yamagata sequences. To generate a dataset of the HA1 region, complete HA sequences were cut at nucleotide position 1030. For each dataset, FULL HA or HA1, we developed a specific classification model. For each model, a subset (training dataset) of approximately 80% of the sequences from each subclade was randomly selected and used to train the random forest model (1000 trees). The remaining 20% of the sequences constituted the testing dataset which was used to evaluate the performance of the respective model.

Both models performed very well, with an accuracy of 0.9952 and 0.9931 for FULL HA and HA1 models, respectively. Additionally, the multiclass AUC was 0.9994 in both cases (supporting information Table S2). Correlation heatmaps, metrics tables, precision-recall curves, and other statistics were generated for each model (supporting information File S1).

To use the app, the user only loads the input file, a FASTA file with unaligned influenza HA or HA1 gene segment query sequences, selects one of the models according to the length of the query sequences (FULL HA or HA1), and presses the run button (Figure 1). To obtain the most accurate results, we recommend using sequences with a proportion of N bases <1%. Since the HA gene allows for more accurate predictions for subtyping based on phylogeny or machine learning models, the other seven influenza genomic segments were not considered in this version but could be incorporated in the future.

Due to the increasing number of laboratories and researchers using sequencing technologies applied to molecular epidemiology, there is an increasing need of easier and faster applications that allows an accurate and specific classification of viral sequences with no need for specialized training. This is particularly relevant for respiratory pathogens such as influenza viruses that cause annual epidemics with up to 60 million ARI cases worldwide and require a continuous monitoring of circulating strains, which is why we believe INFINITy can help researchers working on this area.

ACKNOWLEDGMENTS

We gratefully acknowledge all the authors, the originating laboratories responsible for obtaining the specimens, and the submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We also thank Dr. Andrés Culasso for technical assistance, Dr. Osvaldo Uez for motivation, and Dr. Laura Mojsiejczuk for critical review of the manuscript. We also thank the Centro de Investigación, Docencia y Extensión en Tecnologías de la Información y la Comunicación (CIDETIC, http://cidetic.unlu.edu.ar/), for providing the human and computational resources necessary for this project.

AUTHOR CONTRIBUTIONS

Marco Cacciabue: Formal analysis; methodology; software; validation; visualization; writing-review and editing. **Débora N. Marcone**: conceptualization, investigation, data curation, supervision, validation, visualization, writing – original draft preparation, review & editing.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

FUNDING INFORMATION

This work was supported by a grant from the Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), Argentina (PICT 2018-03603).

PERMISSION TO REPRODUCE MATERIAL FROM OTHER SOURCES

All data are available at GISAID Influenza database.

DATA AVAILABILITY STATEMENT

The application code and instructions are available via Github (https://github.com/marcocacciabue/infinity). Additionally, the web application is available without installation (https://infinity.unlu.edu.ar/).

Marco Cacciabue[1,2,3]
Débora N. Marcone[2,4,5]

[1]*Instituto de Agrobiotecnología y Biología Molecular (IABIMO), Instituto Nacional de Tecnología Agropecuaria (INTA), Hurlingham, Argentina*
[2]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina*
[3]*Departamento de Ciencias Básicas, Universidad Nacional de Luján, Luján, Argentina*
[4]*Cátedra de Virología, Instituto de Bacteriología y Virología Molecular (IBaViM), Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Buenos Aires, Argentina*
[5]*Cátedra de Microbiología, Parasitología y Virología, Facultad de Ciencias Médicas, Pontificia Universidad Católica Argentina, Buenos Aires, Argentina*

Correspondence
Marco Cacciabue, Instituto de Agrobiotecnología y Biología Molecular (IABIMO), Instituto Nacional de Tecnología Agropecuaria (INTA), De los Reseros y N. Repetto s/n, Hurlingham B1686IGC, Buenos Aires, Argentina.
Email: cacciabue.marco@inta.gob.ar; marcocacciabue@yahoo.com

Débora N. Marcone, Cátedra de Virología, Instituto de Bacteriología y Virología Molecular (IBaViM), Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 954, C1113AAD, Buenos Aires, Argentina.
Email: deboramarcone@hotmail.com

ORCID

Débora N. Marcone https://orcid.org/0000-0003-4407-7280

REFERENCES

1. Wang X, Li Y, O'Brien KL, et al. Global burden of respiratory infections associated with seasonal influenza in children under 5 years in 2018: a systematic review and modelling study. *Lancet Glob Health*. 2020;8(4):e497-e510. doi:10.1016/S2214-109X(19)30545-5
2. Lafond KE, Porter RM, Whaley MJ, et al. Global burden of influenza-associated lower respiratory tract infections and hospitalizations among adults: a systematic review and meta-analysis. *PLoS Med*. 2021;18(3):e1003550. doi:10.1371/journal.pmed.1003550
3. WHO. Influenza. Accessed July 22, 2022. https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/vaccine-standardization/influenza
4. Petrova VN, Russell CA. The evolution of seasonal influenza viruses. *Nat Rev Microbiol*. 2018;16(1):47-60. doi:10.1038/nrmicro.2017.118
5. WHO. Candidate vaccine viruses. Accessed July 22, 2022. https://www.who.int/teams/global-influenza-programme/vaccines/who-recommendations/candidate-vaccine-viruses
6. Nextclade. Accessed July 22, 2022. https://clades.nextstrain.org
7. BII. Flusurver - Prepared for the next wave. Accessed July 22, 2022. https://flusurver.bii.a-star.edu.sg/
8. Eisler D, Fornika D, Tindale LC, et al. Influenza classification suite: an automated galaxy workflow for rapid influenza sequence analysis. *Influenza Other Respi Viruses*. 2020;14(3):358-362. doi:10.1111/irv.12722
9. Cacciabue M, Aguilera P, Gismondi MI, Taboga O. Covidex: an ultra-fast and accurate tool for SARS-CoV-2 subtyping. *Infect Genet Evol*. 2022;99:105261. doi:10.1016/j.meegid.2022.105261
10. Humayun F, Khan F, Fawad N, et al. Computational method for classification of avian influenza a virus using DNA sequence information and physicochemical properties. *Front Genet*. 2021;12:599321. doi:10.3389/fgene.2021.599321
11. Borkenhagen LK, Allen MW, Runstadler JA. Influenza virus genotype to phenotype predictions through machine learning: a systematic review. *Emerg Microb Infect*. 2021;10(1):1896-1907. doi:10.1080/22221751.2021.1978824
12. Wang Y, Bao J, Du J, Li Y. Rapid Detection and Prediction of Influenza A Subtype using Deep Convolutional Neural Network based Ensemble Learning. In: *Proceedings of the 2020 10th International Conference on Bioscience, Biochemistry and Bioinformatics*. Kyoto Japan: ACM; 2020:47-51.
13. Wilkinson S. Kmer: An R package for fast alignment-free clustering of biological sequences. 2018.
14. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*. 2017;77:1-17.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.