Predicting virus-host interactions at the species level using machine-learning and a curated viral host range dataset

Eric G. Bastien, Rachel Cable, Cecelia Batterbee, Jacob Evans and Melissa B. Duhaime

Department of Ecology and Evolutionary Biology, University of Michigan, USA

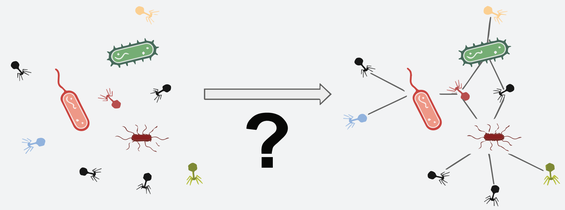
MOTIVATION

The advent of sequencing technology has led to the discovery of hundreds of thousands uncultivated viruses, shedding light on how viruses influence microbial communities and contributes to Earth's biogeochemical processes. Yet, through this process the most important information about any novel virus is lost: who are its hosts?

To remedy this knowledge gap, many tools have been developed to predict the most likely taxa a virus can infect with varying degree of accuracy across taxonomic levels. In addition, comparing those models have been difficult as they all rely on virus-host associations generated from the NCBI database. Finally, there is also a need to predict the complete host range of a virus, which may span across multiple species, sometimes belonging to widely different taxa.

To that end, we collected the largest host range dataset (VHRnet, see Approach below) to answer the following questions:

- Are signals of virus-host coevolution extracted from genomes measurable and significant at the species level?
- How do published predictions tools perform on unseen datasets across different taxonomic levels?
- Can virus-host interactions at the species level be accurately predicted?



APPROACH

VHRnet is a dataset of 8849 lab-verified virus-host interactions at the species taxonomic level collected and digitized from NCBI and literature. This dataset contains both positive data (infection) and negative data (non-infection). This data was at the core to answer each of the questions mentioned above.

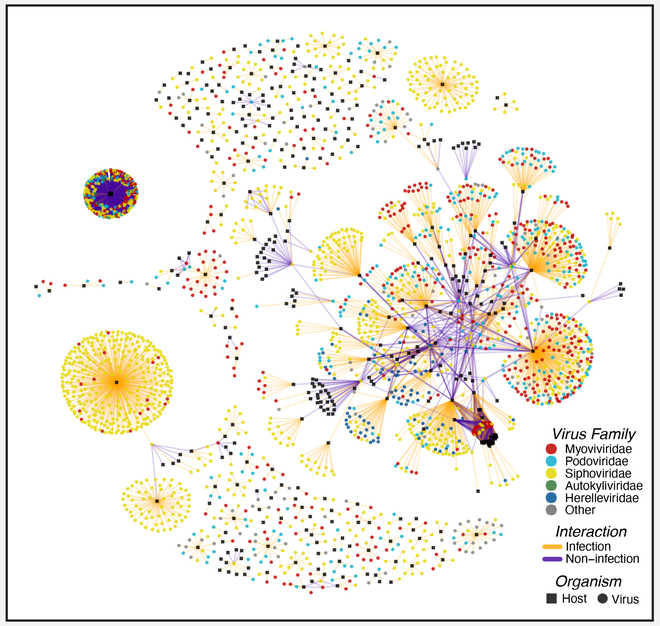


Fig 1. *8849 collected lab-verified virus-host interactions. 2770 of those interactions are* ***infections*** *versus 6079* ***non-infections****. In this network, 2331 unique viruses and 392 unique bacterial species are represented.*

RESULTS

Virus-host coevolution leads to discernible signals in genomes

Commonly used signals of coevolution between virus-host pairs were assessed by comparing infection and non-infection events using the VRHnet dataset.

We confirm that k-mer distances that uses k-length 3 or 6 with the d2* distance metric do encode a discernible signal even at the species level. At k-length of 9, we observe odd behavior where the smallest distances between virus-host pairs are non-infection. Finally, we found that viruses have a significant tendency to remain AT-rich relative to their host by around ~4% in average.

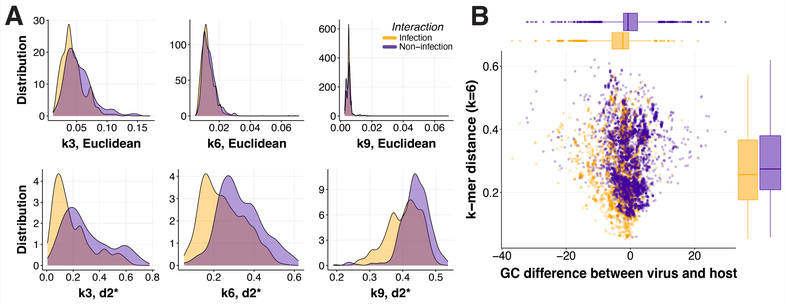


Fig 2. *(A) Comparison between euclidean distance metric and d2* (across), and k-mer length (columns) colored by interaction. (B) A pattern emerges where viruses have a tendency to remain AT-rich while also minimizing k-mer distance respective to their host. Bar plot on top shows variation for GC content, bar plot on the right shows variation for k-mer distance.*

Public databases biases impact predictions

Existing virus-host prediction models rely on NCBI data to train and test their models, thus making a fair comparison between the tools challenging. Complete host range data (part of VHRnet) allows for an unbias comparison between models.

Current models tend to perform well on taxa that those models have seen during training and/or testing, but performs worse on environmental data.

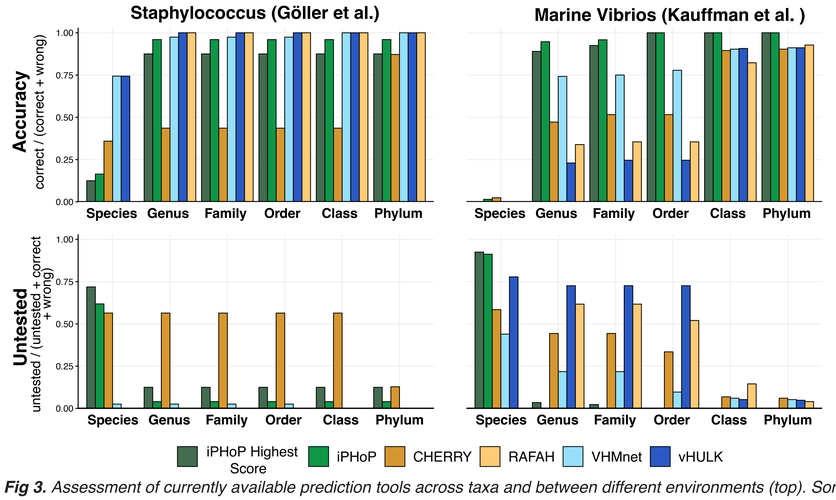


Fig 3. *Assessment of currently available prediction tools across taxa and between different environments (top). Some predictions were untested (bottom), and untested predictions were more prevalent with the environmental habitat.*

Machine learning approach to predict virus-host interactions at the species level

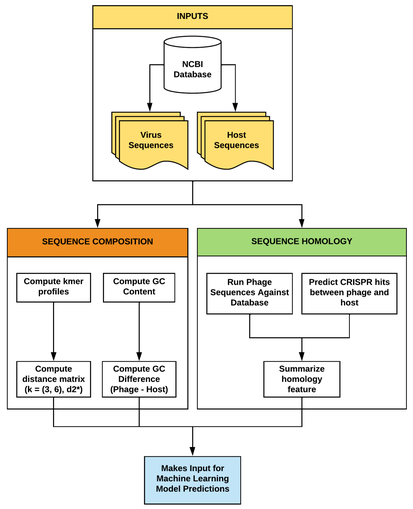


Fig 3. *Workflow in the input features used to train and test the machine learning model to resolve virus-host associations at the species level*

A gradient boosting classifier (using the scikit-learn module in Python) was trained and tested using the VHRnet dataset (70% training, 30% testing) to predict interactions between virus-host pairs at the species taxonomic level.

Features used include k-mer distances, GC difference between virus and host, and homology hits (blastn hits and/or spacers match).

Settings for tree depth and learning rate were tuned by performing 100 grid searches. Bias effect was minimized by bootstrapping (i=100) where the testing and training sets were randomly selected on each iteration of the grid search.

The resulting model has an 87% percent accuracy rate at predicting virus-host interactions at the species level

ROC AUC Score: 0.91 **F1 Score:** 0.79 **Matthew's Correlation:** 0.71

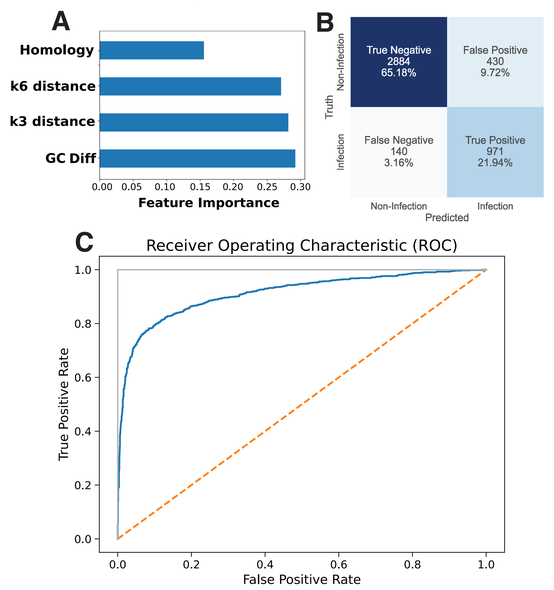


Fig 4. *(A) Feature importance for virus-host prediction based on machine learning model. (B) Confusion matrix. (C) ROC curve of the gradient boosting classifier*

TAKEAWAYS

By compiling lab-verified host range data (VHRnet), we confirmed that signals of virus-host coevolution remain detectable at the species level and can be leveraged for virus-host predictions using machine learning model approaches. This data also allowed for an un-bias comparison between existing prediction models. This study also highlights the importance of experimentally determined host range, and its applicability to further improve virus-host interactions predictions.

Citations

1. Göller, et al. November 2021. Nat. Comm.
2. Kauffman, et al. February 2018. Nature
3. Roux, et al. July 2022. bioRxiv.
4. Shang, et al. May 2022. Brief in Bioinf.
5. Couttinho, et al. July 2021. Patterns.
6. Wang, et al. June 2020. NAR Genomics and Bioinformatics 2
7. Amgarten, et al. August 2022. PHAGE