*Bioinformatics*, 35(19), 2019, 3867–3869
doi: 10.1093/bioinformatics/btz151
Advance Access Publication Date: 1 March 2019
Applications Note

OXFORD

Data and text mining

# VHost-Classifier: virus-host classification using natural language processing

**Ezra Kitson[1] and Curtis A. Suttle[1,2,3,4,*]**

[1]Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada, [2]Department of Earth, Ocean and Atmospheric Sciences, [3]Department of Botany and [4]Institute for the Oceans and Fisheries, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 16, 2018; revised on February 4, 2019; editorial decision on February 24, 2019; accepted on February 26, 2019

## Abstract

**Motivation:** When analyzing viral metagenomic sequences, it is often desired to filter the results of a BLAST analysis by the host species of the virus. VHost-Classifier automates this procedure using a natural language processing algorithm written in Python 3, which takes a list of taxonomic identifiers (taxids) returned from a BLAST query using viral sequences as input. The taxid output is binned by the evolutionary lineage of their host, based on string matching the words in their English names. If VHost-Classifier cannot identify a host, it attempts to bin the sequences by the environment from which the sample originated. VHost-Classifier predicts the evolutionary lineage of the host from the virus name and does not rely on referencing taxids against a database; therefore, it is not constrained by the size of a database and can host classify newly characterized viruses.
**Results:** Benchmarked on a test dataset of 1000 randomly selected viral taxids on the NCBI taxonomy database, VHost-Classifier assigned, with 100% accuracy, a host to the rank of Class for >93% of viruses, and to the rank of Family for >37% of viruses.
**Availability and implementation:** For more information about VHost-Classifier as well as implementation instructions, visit https://github.com/Kzra/VHost-Classifier.
**Contact:** suttle@science.ubc.ca
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

BLAST analysis of viral metagenomic data against other databases often returns many matches (i.e. hits) to other sequences from a diverse range of viruses; however, a researcher may only be interested in viruses that infect specific hosts. In order to filter the results to return only those sequences associated with viruses infecting specific hosts, the analysis must manually be filtered using string lookups, or referenced against a database that has curated virus-host information, such as the Virus-Host DB (Mihara *et al.*, 2016). Manually filtering the results is untenable given that metagenomic runs produce hundreds of thousands of virus reads. Filtering the results using the Virus-Host DB is more viable, however as the database contains <10% of viral sequences on NCBI, many hits from the analysis will not have a host assigned within the database.

Here we present VHost-Classifier, a natural language processing algorithm to automate virus-host classification on a list of viral taxon IDs (vtaxids) that are assigned to sequences during a BLAST search. It groups vtaxids based on the evolutionary lineage of their host, and is therefore useful for filtering vtaxids, or for giving an overview of host diversity for viruses found in a BLAST search, by listing hosts of related viruses.

## 2 Approach

VHost-Classifier is written in Python 3 and requires pre-installation of the ETE3 toolkit to run (Huerta-Cepas *et al.*, 2016). In the default behavior it takes a list of Taxon IDs as input (one Taxon ID per row), which can be extracted as a column from the output of a

BLAST analysis. For more information visit: https://github.com/Kzra/VHost-Classifier.

The pipeline used by the VHost-Classifier algorithm (Supplementary Fig. S1) first references the taxid against the Virus-Host DB and if a host taxon is found it is taken directly from the database. If not found, the taxid is converted to its English name and checked to make sure it is a virus, based on whether it contains a virus descriptive string (e.g. 'Virus', 'Phage' and 'Satellite') in its name. If it is a virus the English name is parsed against a database of common animal names, and if any common animals' names are found they are converted to scientific names that can be used as look-up strings against the NCBI taxonomy database.

Once the conversion is done, the various words in the virus name are parsed to identify the most likely look-up string that identifies the host. To achieve high accuracy, several rules are followed, as detailed below.

First, the viral prefix (-noro, -mega etc.) is ignored, as it is might be confused for the name of a host taxon (e.g. '-noro' may be confused for '*Noronhia*', a genus of flowering plant).

Second, if a string is a genus name, it is concatenated with the following string to make genus and species a single string. This is because species and genus names are confounded across the tree of life, so must be used together in order to be a unique look-up string (e.g. if the virus name is '*Prunella vulgaris* virus 1', the string searched is '*Prunella vulgaris*', as using either the genus or species name alone is ambiguous; both describe multiple taxa in different evolutionary lineages).

Third, if the virus is a strain of Influenza or Norovirus (>75% of all published virus sequences belong to these two taxa), the strain information is also parsed to identify a host. If a specific host cannot be identified for these viruses, because the words in the name do not contain useful host information, it is set as default to be Mammals or Aves depending on the virus subtype (e.g. Influenza A subtypes without strain information are assigned to Aves whilst Influenza B, C, D are assigned to Mammals).

Finally, when choosing between multiple valid look-up strings, basic conventions of English are followed. For example, 'Elephant seal virus' is assigned to a seal not an elephant, as in this case 'Elephant' is acting as an adjective and the virus infects a seal.

Once a look-up string is identified it is used to reference the NCBI taxonomy database, using the ETE3 python toolkit (Huerta-Cepas *et al.*, 2016). The evolutionary lineage of the host is parsed and used to bin the input taxids into a directory tree resembling a phylogenetic tree (Supplementary Fig. S2). Each directory contains csv files that contain the taxids belonging to a particular taxon, and the index positions of these taxids in the original input file. There is also a Counts.csv file that gives the numbers of vtaxids assigned to each taxon within that rank. By default, hosts are binned to the ranks Phylum, Class and Order but this can be set by the user to Phylum, Order and Family.

If a host cannot be assigned to a virus, the taxid is referenced against a customized version of the IMG/VR database (containing only metagenomic and isolate viral sequences) and a set of if-statement rules in order to predict the environment it was sequenced from (Paez-Espino *et al.*, 2017). These taxids are binned according to environment in a separate lineage of the directory tree.

## 3 Performance

When tested and manually checked on a subsample of 1000 vtaxids chosen at random using the Python 'random' module, from the 191 408 vtaxids present on the NCBI taxonomy database (downloaded June 2018), the program assigned a host to >95% of viruses with an overall accuracy of 100%. Over 90% of hosts were resolved to the rank of Class, and 37% could be resolved to the rank of Family (Table 1). Coverage of the subsample dropped sharply when assigning a host to the rank of Order or Family; this is because many viruses on NCBI taxonomy are strains of Influenza or Norovirus that do not have host information in the name. The software assigns a Class to these viruses depending on their subtype (e.g. Influenza A is assigned to Aves) but cannot resolve the host any further.

**Table 1.** Benchmarking accuracy and thoroughness on 1000 randomly selected viral taxonIDs

| Assignation | Accuracy (%) | Coverage (%) | Recall (%) |
|---|---|---|---|
| Superkingdom | 100 | 95.3 | 99.1 |
| Phylum | 100 | 95.3 | 99.7 |
| Class | 100 | 93.2 | 99.7 |
| Order | 100 | 45.5 | 84.0 |
| Family | 100 | 37.4 | 90.7 |

*Note*: 'Accuracy' corresponds to the placement of a vtaxonID in the correct taxon corresponding to rank. 'Coverage' corresponds to the percentage of vtaxids for which an assignment was made. 'Recall' is a measure of the number of vtaxids the authors could assign a host for, based on the virus name, when the software could not (if 100% the software did not overlook any vtaxids).

The recall score reflects the percentage of vtaxids to which the authors could assign a host when the software could not. For each rank assignment the recall score was substantially higher than the coverage, demonstrating that in most cases when the software couldn't assign a host to a vtaxid, it was because the vtaxid did not have enough informative host information in its virus name, and not because the software overlooked useful information.

Based on the virus names, of the 1000 vtaxids queried the authors could assign a host to nine of the 47 viruses, when the software could not assign a host or habitat (Supplementary Table S1). In these cases, the virus names either contained common names of animals, not present in the Common_to_Sci conversion database (e.g. 'threespine stickleback iridovirus') or unusual characters in the name string (e.g. 'cyclovirus ng_chicken 3). A full list of the virus taxids in the subsample and their names can be found on the GitHub page.

The software was able to resolve a host to the rank of Class for 93% of the 191, 408 vtaxids in the NCBI taxonomy database in under 3 h using our lab server (2x Intel Xeon 2 GHz, 32 cores, 512 GB RAM). It is worth noting that there is a strong bias in the hosts of published viruses toward viruses of Chordates (>90% of viruses published on NCBI), and this was reflected in the subsample.

To demonstrate the ability of the software to classify hosts from a real-world metagenomic study, we ran the software on the taxids of ~30k viruses returned from a BLAST analysis of the DNA fraction generated by the Watershed Metagenome Project (Uyaguari-Diaz *et al.*, 2016). It is typical in viral metagenomic studies for many sequences to be assigned as uncharacterized viruses or environmental sequences, without discernible host information in the name. Therefore, in this analysis the coverage scores were lower: 52.6, 51.4, 46, 51.8 and 50.8% for Superkingdom, Phylum, Class, Order and Family host assignments respectively. In this case, there is an increase in coverage from Class to Order because some hosts lack taxonomic information for Class, but have it for Order.

## 4 Conclusions

VHost-Classifier is a tool that will facilitate researchers working on viral metagenomic data by enabling fast filtering of the output from a BLAST search by virus host. In addition, it can give broad insight into the composition of viruses in an environment by host, enabling new and interesting questions to be asked during viral metagenomic research.

## Funding

This work was supported by a Discovery Grant [2015-05896] from the Natural Sciences and Engineering Research Council of Canada.

*Conflict of Interest*: none declared.

## Acknowledgements

We would like to thank Phillipe Tortell, Steven Hallam, Christoph Deeg, Marli Vlok, Gideon Mordecai, Jan Finke, Jessica Caleta and Henry Laney for their encouragement and guidance in creating and testing the program.

## References

Huerta-Cepas,J. *et al*. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol*., **33**, 1635–1638.

Mihara,T. *et al*. (2016) Linking virus genomes with host taxonomy. *Viruses*, **8**, 66.

Paez-Espino,D. *et al*. (2017) IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res*., **45**, D457–D465.

Uyaguari-Diaz,M.I. *et al*. (2016) A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome*, **4**, 20.