Title: An Integrated Machine-Learning Framework for Reliable Host Prediction of Uncultivated Phages

Authors: Simon Roux[1] (sroux@lbl.gov), Andrew Tritt[2]

Institutions: [1]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA

Project Goals: Environmental viral diversity is quickly being mapped through large-scale metagenomics (meta)analyses. A major challenge of such approach, compared to traditional viral isolation, is the lack of host information for uncultivated viruses. Over the last several years, multiple tools have been released to predict host taxonomy from (partial) genome sequences of uncultivated viruses. Here, we review the major types of host predictions proposed so far, and describe the prototype of a new neural network framework able to integrate results from multiple tools in a single reliable host genus prediction.

Abstract text: Viruses are critical components of soil microbial ecosystems. By shaping microbial communities' structure and altering host cell metabolism during infection, viruses exert strong constraints on microbiomes, with downstream effects on nutrient cycling and metabolic outputs. Viral genomic diversity in the environment, especially in soil, is progressively being mapped primarily through the assembly of novel viral genomes from metagenomes. In the last five years alone, the number of such "uncultivated virus genomes" available in public databases has increased by more than 3 orders of magnitudes (from a few hundreds to several millions[1]). These genomes have enabled multiple discoveries on the diversity and distribution of viruses across different ecosystems, yet one limitation inherent to these datasets is the lack of host information for these viruses.

Linking novel uncultivated viruses to their host(s) is a critical step towards understanding the influence and potential impacts of these viruses on microbiomes and ecosystems. Accordingly, more than 15 different tools have been released over the last few years aiming at predicting virus:host pairs and/or host taxonomy for uncultivated viruses. Overall, these rely on four major types of genomic signal: (i) sequence similarity to known viruses, (ii) sequence similarity to putative host genomes, including due to host-encoded CRISPR-Cas systems, horizontal gene transfer, and provirus integration, and (iii) similarity in terms of nucleotide composition (i.e., k-mer frequency) between virus and host genome[2-5]. These different approaches differ greatly in their recall (i.e., ability to predict host taxonomy for as many viruses as possible) and false-discovery rate (i.e., frequency at which the predicted taxonomy correspond to the correct host), so that aggregating the results obtained from different tools on a single virus is challenging[6-7].

We describe here the prototype of a machine-learning framework taking as input multiple results of host prediction tools and using a deep neural network structure to provide as output a single host taxonomy prediction, at the genus rank, along with a confidence score. We illustrate how

this integration step maximizes both recall and precision, enabling robust host prediction for more input sequences than any individual tool. In addition, rather than directly predicting a host taxonomy, the neural network is designed to learn to distinguish patterns of "reliable" and "unreliable" host prediction based on a combination of all signals considered. Hence, it is not limited by the virus-host pairs currently described, and could be applied to entirely novel virus and host communities.

Overall, integrating independent signals for host prediction appears to be a promising approach to progressively populate uncultivated virus genome databases with reliable host information at a satisfactory taxonomic resolution (i.e., genus rank). Coupled with a continuing expansion of the global collection of isolated phages and innovative experimental methods to link uncultivated viruses to hosts, these pave the way towards a more comprehensive reconstruction of virus:host network from complex microbial communities.

References/Publications

1. Roux, Simon, David Páez-Espino, I-Min A Chen, Krishna Palaniappan, Anna Ratner, Ken Chu, T B K Reddy, et al. 2020. "IMG/VR v3: An Integrated Ecological and Evolutionary Framework for Interrogating Genomes of Uncultivated Viruses." Nucleic Acids Research, 1–12. https://doi.org/10.1093/nar/gkaa946.
2. Edwards, Robert A., Katelyn McNair, Karoline Faust, Jeroen Raes, and Bas E. Dutilh. 2016. "Computational Approaches to Predict Bacteriophage-Host Relationships." FEMS Microbiology Reviews 40 (2): 258–72. https://doi.org/10.1093/femsre/fuv048.
3. Galiez, Clovis, Matthias Siebert, François Enault, Jonathan Vincent, and Johannes Söding. 2017. "WIsH: Who Is the Host? Predicting Prokaryotic Hosts from Metagenomic Phage Contigs." Bioinformatics 33 (19): 3113–14. https://doi.org/10.1093/bioinformatics/btx383.
4. Ahlgren, NA, Jie Ren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. 2016. "Alignment-Free d2 * Oligonucleotide Frequency Dissimilarity Measure Improves Prediction of Hosts from Metagenomically-Derived Viral Sequences." Nucleic Acids Research 45 (1): 39–53. https://doi.org/10.1093/nar/gkw1002.
5. Zhang, Ruoshi, Milot Mirdita, Eli Levy Karin, Clovis Norroy, Clovis Galiez, and Johannes Soeding. 2020. "SpacePHARER: Sensitive Identification of Phages from CRISPR Spacers in Prokaryotic Hosts." BioRxiv, 2020.05.15.090266. https://doi.org/10.1101/2020.05.15.090266.
6. Wang, Weili, Jie Ren, Kujin Tang, Emily Dart, Julio Cesar Ignacio-Espinoza, Jed A Fuhrman, Jonathan Braun, Fengzhu Sun, and Nathan A Ahlgren. 2020. "A Network-Based Integrated Framework for Predicting Virus–Prokaryote Interactions." NAR Genomics and Bioinformatics 2 (2): 1–19. https://doi.org/10.1093/nargab/lqaa044.
7. Zhang, Fan, Fengxia Zhou, Rui Gan, Chunyan Ren, Yuqiang Jia, Ling Yu, and Zhiwei Huang. 2019. "PHISDetector: A Web Tool to Detect Diverse in Silico Phage-Host Interaction Signals." BioRxiv. https://doi.org/10.1101/661074.

Funding statement: This work was supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research, Early Career Research Program awarded under UC-DOE Prime Contract DE-AC02-05CH11231. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.