Accurate Identification of Hosts from Environmental Viruses Using Deep Learning Networks and High Level Features

Deyvid Amgarten, Bruno Iha, Aline Maria da Silva, João Carlos Setubal

Laboratório de Técnicas Especiais, Hospital Israelita Albert Einstein

Abstract

Microbial genomics has been experiencing expressive changes in the last decade, mostly due to improvements in environmental sampling and sequencing techniques generally known as metagenomics. Thousands of new complete virus genomes are made available every year, but experimental characterization has not kept pace. In particular, information about the host of viruses whose genomes have been sequenced is commonly lacking. This is one of the most essential information needed for cultivation, isolation and many other microbiological characterization techniques. Here we present a toolkit called vHULK (Viral Host Unveiling Kit), which predicts taxonomic and biological attributes of a virus' host. Our tool receives complete or high quality virus draft genomes as input, and provides as output: 1. Probability scores of predicted host genus and/or species; 2. Tables of cross-probabilities among possible hosts; and 3. information theory measurements and a rank of informative features about virus-host relationships. Our methodology is based on feature extraction of virus' genomes with known hosts leading to matrices containing thousands of features. After feature extraction and feature normalization, matrices were used to train a multilayer perceptron deep neural network classifier. Performance was measured by assessing validation and training curves, as well as by measurement in batch test sets. For a multiclass problem of 62 possible bacterial host genus, vHULK presented an average accuracy of 98% in the test set. When development is finished, vHULK will be freely available through user-friendly python scripts at a Github repository.

Funding: This work had the support from CAPES, CNPq and FAPESP research funding agencies